UNIVERSITÄT ZU LÜBECK

Aus der Klinik für Psychiatrie und Psychotherapie
der Universität zu Lübeck
Direktor: Prof. Dr. Stefan Borgwardt

Understanding (dys)functional self-belief formation: Evidence from experimental, computational, and clinical neuroscience

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

Aus der Sektion Naturwissenschaften

vorgelegt von
Nora Czekalla
aus Hamburg

Lübeck, 2024

1. Berichterstatter: Prof. Dr. Sören Krach
2. Berichterstatter: Prof. Dr. Mario Gollwitzer

Tag der mündlichen Prüfung: 22. April 2025

Zum Druck genehmigt. Lübeck, 06. November 2025

Contents

Abstract

Self-beliefs are essential to our identity and emotional well-being. They are formed through self-related feedback from the environment. The processing of feedback is influenced by self-related motivations and emotional experiences. This can lead to biased weighting of incoming information when updating beliefs depending on whether it is better or worse than expected. In depression, these biases can perpetuate maladaptive beliefs. This dissertation explores how novel self-beliefs are formed, using neuroscientific and computational methods to model belief formation in response to feedback. Across four studies, biases in self-belief formation were examined in individuals with and without depression, considering individual and contextual factors such as affective experiences, symptom burden, and stress. Participants engaged in a learning task with trial-by-trial performance feedback in relatively unfamiliar domains, designed to elicit the formation of novel self-beliefs.

Self-belief updates were best described by prediction error valence, with consistent findings of a negativity bias specific to self-beliefs but absent for beliefs about others. Study 1 investigated audience effects on self-belief formation and links to global prior self-beliefs indicated by self-esteem. It showed that more negative global self-beliefs were associated with more negative self-belief formation. In a public context, belief formation was more negative in those with higher subclinical social anxiety. Study 2 combined functional magnetic resonance imaging (fMRI), pupillometry, and emotion ratings to demonstrate that more negative self-belief formation was associated with the experience of more embarrassment and less pride, as well as more arousal, as indicated by pupil dilation. Neurally, it was linked to a heightened activity to more negative prediction errors in the anterior insula, amygdala, midbrain regions, and medial prefrontal cortex. Study 3 examined individuals with depression and healthy controls using fMRI. Individuals with depression showed stronger reactivity in the insula to negative compared to positive prediction errors. Despite no group differences in behavior, individuals with a higher symptom burden exhibited more biased belief formation, displaying more negativity towards themselves and more positivity towards others. In individuals with depression, in particular, symptom burden was linked to reduced updates following positive prediction errors. Study 4 showed that self-belief formation was less negatively biased following social-evaluative threat, which was associated with better recovery from stress-induced negative affect.

In summary, the studies show a negativity bias when forming beliefs about one's ability. This may be related to motivations present in performance contexts like improvement motivation or motivation to avoid failure with a threat driven focus on negative feedback. The findings also highlight the critical role of individual factors like current affective experience, global prior self-beliefs, and symptom burden, alongside contextual factors like social evaluative stress in self-belief formation. Neurally, the studies identify the insula as central to processing affective, self-related feedback linked to biased updating. Notably, in individuals with depression, reduced self-belief updating following positive prediction errors with increased symptom burden suggests a diminished receptivity to corrective positive feedback, which may perpetuate maladaptive self-beliefs.

Zusammenfassung

Selbstüberzeugungen sind wesentlich für unsere Identität und unser Wohlbefinden. Sie werden durch selbstbezogene Rückmeldungen aus der Umwelt gebildet. Die Verarbeitung dieser Rückmeldungen wird von selbstbezogenen Motivationen und emotionalem Erleben beeinflusst. Dies kann zu einer verzerrten Gewichtung eingehender Information führen, je nachdem, ob sie besser oder schlechter als erwartet ist. Bei Depressionen kann diese verzerrte Informationsverarbeitung maladaptive Überzeugungen aufrechterhalten. In dieser Dissertation wird untersucht, wie neue Selbstüberzeugungen gebildet werden, wobei neurowissenschaftliche und computergestützte Methoden zur Modellierung der Überzeugungsbildung als Reaktion auf Feedback eingesetzt werden. In vier Studien wurden Verzerrungen bei der Bildung von Selbstüberzeugungen bei Personen mit und ohne Depression untersucht, wobei individuelle und kontextuelle Faktoren wie affektives Erleben, Symptombelastung und Stress berücksichtigt wurden. In mehreren Durchgängen bearbeiteten die Teilnehmenden eine Lernaufgabe mit Leistungsfeedback bezüglich relativ ungewohnter Domänen, was die Bildung neuer Selbstüberzeugungen anregen sollte.

Die Anpassung der Selbstüberzeugung an Feedback wurden am besten durch die Vorhersagefehlervalenz beschrieben, mit dem konsistenten Befund einer Negativitätsverzerrung. Diese war für Selbstüberzeugungen spezifisch und bei Überzeugungen über andere nicht vorhanden. Studie 1 untersuchte die Auswirkungen von Publikum auf die Bildung von Selbstüberzeugungen sowie den Zusammenhang mit globalen Vorannahmen über sich, erfasst durch den Selbstwert. Es zeigte sich, dass negativere globale Vorannahmen mit einer negativeren Selbstüberzeugungsbildung verbunden sind. In einem öffentlichen Kontext war die Überzeugungsbildung bei Personen mit höherer subklinischer sozialer Ängstlichkeit negativer. In Studie 2 wurden funktionelle Magnetresonanztomographie (fMRI), Pupillometrie und Emotionsratings kombiniert, um zu zeigen, dass die Bildung negativer Selbstüberzeugungen mit dem Erleben von mehr Peinlichkeit und weniger Stolz sowie mit erhöhter Anspannung verbunden war, was durch die Pupillenerweiterung gemessen wurde. Auf neuronaler Ebene wurde dies mit einer erhöhten Aktivität bei negativen Vorhersagefehlern in der anterioren Insula, der Amygdala, Regionen im Mittelhirn und dem medialen präfrontalen Kortex in Verbindung gebracht. In Studie 3 wurden Personen mit Depressionen und gesunde Kontrollpersonen mittels fMRI untersucht. Personen mit Depressionen zeigten eine stärkere Reaktivität in der Insula auf negative im Vergleich zu positiven Vorhersagefehlern. Obwohl es keine Gruppenunterschiede im Verhalten gab, wiesen Personen mit einer höheren Symptombelastung eine verzerrtere Überzeugungsbildung auf. Dies drückte sich durch eine stärker negative Überzeugungsanpassung bei Selbstüberzeugungen und stärker positiver Anpassung bei Überzeugungen über andere aus. Spezifisch bei Personen mit Depressionen war die Symptombelastung mit einer geringeren Anpassung nach positiven Vorhersagefehlern verbunden. Studie 4 zeigte, dass die Bildung von

Selbstüberzeugungen nach Stress durch soziale Bewertung weniger negativ verzerrt war, was mit einer besseren Erholung von stressbedingtem negativem Affekt verbunden war.

Zusammenfassend zeigten die Studien eine negative Verzerrung bei der Bildung von Überzeugungen über die eigenen Fähigkeiten auf. Dies kann mit Motivationen in Leistungskontexten zusammenhängen, wie z. B. Verbesserungsmotivation oder Motivation zur Vermeidung von Misserfolgen mit einem bedrohungsgesteuerten Fokus auf negatives Feedback. Die Ergebnisse unterstreichen die entscheidende Rolle individueller Faktoren wie aktuelles affektives Erleben, globale Vorannahmen über sich und Symptombelastung sowie kontextueller Faktoren wie sozialer Stress bei der Bildung von Selbstüberzeugungen. Auf neuronaler Ebene identifizieren die Studien die Insula als zentral für die Verarbeitung affektiver, selbstbezogener Rückmeldungen, die mit einer verzerrten Anpassung von Selbstüberzeugungen an Feedback verbunden sind. Insbesondere bei Menschen mit Depressionen deutet eine geringere Anpassung der Selbstüberzeugung nach positiven Vorhersagefehlern bei erhöhter Symptombelastung auf eine verminderte Empfänglichkeit für korrigierendes positives Feedback hin. Dies trägt möglicherweise zu einer Aufrechterhaltung maladaptiver Selbstüberzeugung bei.

1 General Introduction

1.1 Intro to this thesis

Our beliefs are the lenses through which we see ourselves, others, and the world (Ellis & Dryden, 1997; Yeager & Dweck, 2012). They shape our expectations, guide our attention and perception, and drive our actions (Hughes & Zaki, 2015). This means we are not passive recipients of our environment; rather, we actively generate predictions (Barrett, 2017a; Clark, 2013; Friston, 2010). This shapes how we process information and can result in a biased integration of new information, especially in response to self-related feedback. Predictions are guided by our learned beliefs shaped by personal learning history (Bandura, 1977; Friston, 2005; Wolpert & Miall, 1996). While research has frequently addressed self-beliefs and single belief updates, the learning history underlying self-belief formation has received less attention. This thesis aims to investigate the formation of relatively novel self-belief and explore the biases inherent in this process. Using a newly developed learning paradigm, self-belief formation will be computationally modeled in four consecutive studies. Relationships between biases and neural correlates, individual factors such as self-esteem or affective experience, as well as contextual factors such as stress will be examined. Both healthy participants and individuals with depression will be included to investigate possible psychopathological dynamics in self-belief formation, with a special focus on a dimensional perspective on psychopathology. This should enable the translation into the clinical context to discuss how biased self-beliefs can be addressed.

1.2 Introduction to belief updating

To navigate a changing environment, we refine our learned beliefs to make more precise predictions (Bandura, 2001; Clark, 2013). This means beliefs are dynamic and can evolve as individuals encounter situations that deviate from expectations (Markus & Wurf, 1987; Mokady & Reggev, 2022). This belief updating process results from minimizing prediction errors, the mismatch between prediction and actual outcome (Friston, 2005). Prediction errors can be minimized by either updating the belief towards the new information or by engaging in actions to change the environment to make it match the predictions; for example, reinterpreting incoming information (Mokady & Reggev, 2022). The latter is referred to as active inference and suggests that perception and action are deeply intertwined. As mentioned above, this means that the brain is not a passive observer but an active participant in shaping sensory experiences (Bromberg-Martin & Sharot, 2020; Clark, 2013; Friston, 2010). The more one tends toward active inference in a specific situation, the less deviating information is integrated, making it more likely that a belief remains unchanged.

Whether a belief is updated or not depends on various factors. One could assume that whenever a person encounters new information that deviates from their belief, the belief is adjusted accordingly. However, this is not always the case, as optimizing the accuracy of future predictions and maximizing external outcomes is not the only goal of this process.

Instead, beliefs are also associated with an internal outcome such as pleasant feelings; that is, beliefs have value in themselves (Bromberg-Martin & Sharot, 2020). For example, people like to be right and prefer to hold beliefs with high certainty. This can result in prioritizing certain information and neglecting others (Leong et al., 2019). Attention control, information selection, and belief updating are motivated by optimizing internal outcomes in order to approach positive states and avoid unpleasant states (Loewenstein, 2006).

While also internal outcomes can depend on a belief's accuracy (e.g., accurately positive beliefs can lead to pride and overly positive, i.e., inaccurate beliefs to disappointment once the test result is received), there are also accuracy-independent external or internal outcomes that determine a belief's value (Sharot et al., 2023). For example, adhering to the same beliefs as a social in-group can be associated with social acceptance (external) and a positive sense of belonging (internal). Since most beliefs relate to hidden, not directly observable, aspects of the self or the world, they often involve a degree of uncertainty. Also, incoming information can be vague or precise (Sharot et al., 2023). These levels of uncertainty can also contribute to whether a belief is updated, with less weight given to vague information (Ernst & Banks, 2002; Sharot et al., 2023). In the following, the significance of the valence of beliefs is emphasized, which is particularly relevant for self-beliefs.

Self-belief updating

Self-beliefs refer to the perceptions and attitudes individuals hold about their abilities, traits, values, or life circumstances (Bandura, 1982; Baumeister, 2019; Markus & Nurius, 1986). They go along with a special emotional investment and specific motivational factors (see 1.2.1; Deci & Ryan, 2000). This means the internal outcomes determining a belief's value are essential for self-beliefs. For example, holding a self-serving belief is associated with positive affect and mental health in the long term (Taylor & Brown, 1988). To increase internal outcomes like happiness or pride, prediction errors receive more or less weight in the updating process depending on their valence. Therefore, some self-beliefs and their associated affect can be approached and preferably integrated, and others avoided. A tendency to update some beliefs while engaging in active inference to maintain other beliefs results in an overall biased self-belief updating (Sharot & Garrett, 2016). The link between self-belief updating and affect is recursive as beliefs are associated with certain affects, and certain affective states make certain beliefs especially present. This means affect influences the process of belief updating, and belief updating, in turn, influences the affective state (Bromberg-Martin & Sharot, 2020). As we actively construct our environment through predictions and active inference, the processing of incoming information must be considered against the background of self-related motivation and associated emotions in order to understand biased belief updating (Sharot et al., 2023).

Many studies on healthy samples have reported belief updating in a self-serving way by a stronger adjustment towards information that is better than expected (Eil & Rao, 2010; Korn et al., 2012; Kuzmanovic et al., 2016; Möbius et al., 2011; Sharot et al., 2011). This bias has been referred to as positivity or optimism bias. It has been shown when updating

beliefs about one's future (Kuzmanovic & Rigoux, 2017; Sharot et al., 2011), personality (Korn et al., 2012), academic performance (Villano et al., 2023), intelligence, or appearance (Eil & Rao, 2010; Möbius et al., 2011). Another frequently reported bias in belief updating is the confirmation bias. This means individuals seek and process information in a way that matches their prior beliefs (Swann & Read, 1981b). When facing contrary evidence, individuals might engage in active inference and critically scrutinize it, find errors, or devalue the relevance to maintain their existing belief (Mokady & Reggev, 2022).

Other studies in healthy samples have shown negative biases in updating self-beliefs (Brotzeller & Gollwitzer, 2024; Ertac, 2011; Zamfir & Dayan, 2022). Unlike self-belief updating paradigms with a positive bias, these paradigms typically examine belief updating in a performance context with active task execution. This shows that self-belief updating can be biased in both directions. It implies that the context may be essential for understanding self-belief updating (see 1.3.2), as different learning tasks elicit different self-related motivations and affective experiences (Elliot & McGregor, 2001). This could influence the direction and extent to which learning is biased. Affective states and motivations can also depend on different learning conditions, such as learning under stress (Garrett et al., 2018a; Globig et al., 2022). Thus, the conditions in which a learning task is embedded are also important when studying the dynamics of self-beliefs.

To better understand how exactly self-related motivations and emotions relate to biased self-belief updating, these concepts will first be explained in more detail. Afterward, they will be connected to specific task characteristics, such as the type of self-belief addressed or the type of feedback, as well as to the contextual task embedding. In this way, a framework of contextual factors, corresponding motivations and emotions, and potential biases will be provided. The studies in this thesis will later be integrated into this, which should enable a structured discussion of the findings.

1.3 Drivers of biased self-belief updating

1.3.1 The role of affect and motivation in self-belief updating

Self-related motivations shape how we interact with our environment, seek or avoid certain social feedback, and judge its validity (Hughes & Zaki, 2015). Therefore, motivation and the associated emotions are crucial for understanding self-related feedback processing (Somerville et al., 2010; Yoon et al., 2018). Whether feedback is categorized as useful or false, rewarding or threatening, can influence to which extent it is used to adjust a self-belief (Kluger & DeNisi, 1996; Sharot et al., 2023).

The motivations in feedback processing can be broadly categorized into the approach of rewarding and avoidance of negative outcomes (Elliot et al., 2006; Gable, 2006). On the one hand, individuals seek feedback that confirms or enhances their positive self-views, driven by a desire for a positive feeling about themselves (Hepper et al., 2010; Swann, 1983). This pursuit of positive reinforcement helps maintain self-esteem and cognitive consistency (Abelson et al., 1968; Crocker & Park, 2004; Harter, 1999). On the other hand, individuals engage in threat monitoring to be able to avoid negative social evaluation

and the accompanying negative emotions, such as embarrassment or shame, in the long term (Lewis, 2008; Strachman & Gable, 2006). By looking at these dual motivations and the associated emotions, we can gain insights into their impact on biases in updating self-beliefs.

Self-related approach motivations

Self-related approach motivations revolve around self-enhancement, self-verification, self-improvement (Sedikides & Strube, 1997), and self-assessment (Brown et al., 2015; Leary, 2007). Self-enhancement theory suggests that individuals aim to achieve positive self-evaluations to hold self-serving views about themselves (Leary, 2007; Sedikides & Strube, 1997; Taylor & Brown, 1988). This drive leads to overestimations of one's abilities and positive attributes, also called the above-average effect (Sedikides & Gregg, 2008; Zell et al., 2020). Self-enhancement motivation results in a self-serving information selection, for example, downward social comparisons (Wills, 1981; Wood, 1989) and self-serving interpretations like attributing positive outcomes to oneself and negative outcomes to external factors (Miller & Ross, 1975). Aside from rewarding feelings of personal competence or self-worth, this motivation also has the social component to be positively evaluated by others (Leary, 2007). This can shape social interaction by managing one's impression in front of others, for example, by presenting oneself in a favorable or socially conform way (Leary & Kowalski, 1990). Self-enhancement motivation and impression management are conceptually related to performance-approach goals in educational settings (Ames & Archer, 1988). It refers to the focus on one's performance in comparison to others and on performance measures such as grades with the motivation to be better than others (Ames, 1992). Self-enhancement can lead to distorted predictions about one's abilities, behavior, or life circumstances. For example, people overestimate when they complete work projects, are unrealistically optimistic about their health risks, overestimate the likelihood that they will behave desirably, or how prevalent their opinions and preferences are among their peers (Dunning et al., 2004).

Self-verification theory posits that individuals aim to confirm pre-existing self-beliefs through their environment and interactions (Swann, 1983). They seek information and process it in a way that aligns with their current self-view (Swann & Read, 1981a). This maintains cognitive consistency and reinforces their beliefs about themselves (Abelson et al., 1968). It creates the positive feeling of a coherent and predictable world (Swann, 2012), which corresponds to the need for security, control, and, therefore, autonomy (Deci & Ryan, 2000). On a social level, people strive for others to see them as they see themselves (Swann, 1983). In social interactions, people engage in behaviors that elicit confirmatory feedback and prefer to associate with others who verify their self-views. By aligning others' perceptions with their own, individuals can reduce uncertainty and create a stable social identity (Swann, 1983; Swann et al., 1994).

Self-assessment is the desire to obtain accurate and diagnostically valuable information about oneself (Brown et al., 2015; Leary, 2007). It stems from the need to align the self-perception with reality. By engaging in self-assessment, people aim to reduce uncertainty

about their strengths and weaknesses, which can guide decision-making, for example, in career orientation and support self-improvement (Sedikides & Strube, 1997).

The motivation for **self-improvement** (Sedikides & Hepper, 2009) is regarded here as synonymous with the learning or mastery goal in educational settings (Ames & Archer, 1988) that aims to improve or develop new skills. The emphasis is on learning itself instead of performance compared to others, which is the focus of performance-approach goals/ self-enhancement. For self-improvement, mastery is seen as dependent on effort (Ames & Archer, 1988). Central to this motivation is self-efficacy, which is the belief in one's capacity to improve abilities through effort and learning (Bandura, 1977). Individuals who perceive their abilities as changeable are driven by a desire to gain competence and autonomy (Deci, 2009). This motivation propels individuals to seek out and utilize feedback to enhance their skills and achieve their goals (London, 2003). In contrast to self-enhancement, negative information can be valuable here because it provides information about where effort is needed (Sedikides & Hepper, 2009). However, the type of feedback matters. Negative feedback that is specific and behavior-oriented and does not threaten self-worth too much promotes the approach motivation of self-improvement (Engerer et al., 2019). Upward social comparison feedback is compatible with self-improvement motivation when self-threat is low, and hope for change is induced (e.g., "I can make it, too."; Sedikides & Hepper, 2009). If negative feedback is based on globally devaluing criticism, self-worth is more likely under threat, and avoidance motivation tends to drive feedback processing (Hattie & Timperley, 2007).

Self-related avoidance motivations

While self-related feedback can be rewarding and motivate to approach positive or confirming self-images, it can also threaten self-esteem or social belonging (Hoefler et al., 2015; Somerville et al., 2010). Contexts that primarily create a threat of negative social evaluation or exclusion go along with the motivation to avoid this (Smart Richman & Leary, 2009). Thus, self-related avoidance motivation is also important for understanding feedback processing and self-belief updating. A context that typically elicits the fear of failure is a performance context. Whenever people actively engage in a task and receive direct feedback on their performance, they are motivated to avoid negative evaluation. Avoidance motivation in performance contexts can be distinguished into performance-avoidance goals and mastery-avoidance goals (Elliot & McGregor, 2001). Performance-avoidance goals aim to avoid performing worse than others, while mastery-avoidance goals focus on avoiding a decline in personal competence. An underlying concern is usually being devalued by others. To avoid this, negative feedback contains special information content as an indication of possible social devaluation (Baumeister & Leary, 1995). Thus, attention is shifted to potential threats (Strachman & Gable, 2006), also called threat monitoring (Shechner & Bar-Haim, 2016). To some extent, threat monitoring is adaptive and present in healthy individuals (Shechner & Bar-Haim, 2016). For example, when giving a talk, most people experience increased anxiety and instinctively scan for critical or bored facial expressions in the audience. As humans are social beings who want to avoid social exclusion, monitoring potential negative evaluation is helpful to a certain

extent (Pickett et al., 2004). Especially in the short term, it can regulate anxiety by fostering a sense of control or catching a reassuring audience response. However, prolonged threat monitoring increases the likelihood of detecting negative cues, so it usually maintains or increases anxiety in the long term. To a greater extent, threat monitoring occurs in various mental disorders, especially anxiety disorders (Shechner & Bar-Haim, 2016). Threat monitoring of social feedback is particularly pronounced in social anxiety, where individuals exhibit increased sensitivity to negative evaluations and potential failure (Clark & Wells, 1995; Heimberg et al., 2014). This leads to heightened threat perception, reinforcing avoidance behavior, and increasing inner overrepresentation of one's failures. It results in a vicious cycle of negative self-evaluation and further avoidance (Clark & Wells, 1995). Understanding the dynamics of self-beliefs in social-evaluative contexts that evoke avoidance motivation provides an important bridge for the transfer to clinical contexts.

Self-related motivation and self-conscious emotions

As satisfaction or frustration of needs tied to motivational states is inherently linked to emotions, studying affect is crucial for understanding potential motivations underlying biases in self-belief updating (Leary, 2007). A subset of emotions called self-conscious emotions relates to our sense of self and perceived judgments by others (Lewis, 2008; Müller-Pinzler et al., 2017; Tracy & Robins, 2004). They usually arise in actual or imagined social contexts and involve self-reflection, including evaluating one's action against internal or external standards or norms. They are important for social behavior and self-regulation to promote social acceptability and avoid social exclusion. Especially the emotions of guilt, shame, embarrassment, social anxiety, and pride stem from inferences about others' actual or imagined evaluations (Leary, 2007). Shame arises when individuals perceive that they failed to meet their own or others' expectations. This norm violation is seen as an expression of character as a whole and not a consequence of specific behavior, which would rather trigger guilt (Dearing & Tangney, 2003). Social anxiety and embarrassment relate to others' perceptions and evaluations (Leary, 2007). Social anxiety expresses the fear of negative social evaluation, and embarrassment expresses the perceived or imagined social failure that the individual associates with a negative social evaluation (Schlenker & Leary, 1982; Tangney et al., 1996). Pride arises from perceived responsibility for socially valued outcomes or the belief to be a socially valued person (Stolz et al., 2020; Tracy & Robins, 2007a). Emotions also have a behavioral component. For example, the experience of pride motivates people to pursue something further (Williams & DeSteno, 2008), and embarrassment rather motivates them to discontinue a current behavior or appease others (Apsler, 1975; Feinberg et al., 2012).

Affect and motivation are intertwined. Motivations aim to satisfy basic needs, while emotions serve to alert and evoke corresponding reactions (Jenkins & Oatley, 1996). How self-related information is selected, interpreted, and integrated into the self-concept is driven by self-related motivations to approach an appetitive affective state and avoid an aversive affective state (Dunning et al., 2004; Hughes & Zaki, 2015). The self-related motivations described can thus be associated with the following affective states (although

there is no claim to completeness or exclusivity here). Self-verification can result in the positive affect of consistency and predictability. A stable and coherent self-image reduces cognitive dissonance and promotes a sense of control and security (Morvan & O'Connor, 2017; Ryan & Deci, 2000; Sedikides & Strube, 1997; Swann, 1983). Validation from others can also lead to positive social affects, such as affection or feeling understood and connected (Laurenceau et al., 1998; Swann, 2012). The feeling of having been "right" can also lead to pride (Tracy & Robins, 2007a). Pride in being a valued person can primarily be associated with self-enhancement, for example, when individuals feel competent, successful, or superior (Leary, 2007). Additionally, self-enhancement can be linked to confidence as a reinforcing feeling that strengthens positive self-beliefs (Möbius et al., 2011; Sedikides & Strube, 1997). This strengthening of one's competence and worth can also lead to a feeling of security as it reduces anxiety or self-doubt. A feeling of optimism can also occur as self-enhancement tends to boost positive outlooks for upcoming challenges (Taylor & Brown, 1988). Pride in specific accomplishments can be associated with self-improvement motivation (Tracy & Robins, 2007b). Also, self-improvement can evoke a sense of progress or growth (Sedikides & Strube, 1997). By expanding one's competence, a feeling of self-efficacy and autonomy can also be experienced (Bandura, 1977).

The expectation of negative social evaluation is associated with social anxiety driving self-related avoidance motivation. This can involve, for example, avoiding a social situation from the outset or trying to regulate anxiety with safety behavior or threat monitoring (Clark & Wells, 1995). Embarrassment makes people avoid situations where they could expose themselves to ridicule or a feeling of failure (Lewis, 2008; Miller, 1996; Müller-Pinzler et al., 2015). When one tends to feel inadequate, flawed, or unworthy, shame can lead to avoidance of situations that may expose these perceived flaws. Also, guilt can drive avoidance when a person feels responsible for wrongdoing and wishes to avoid facing the consequences or reminders of their actions (Dearing & Tangney, 2003). Other emotions associated with negative social evaluation, punishment, exclusion, and, thus, avoidance motivation include frustration, helplessness, humiliation, self-doubt, or anger (Smart Richman & Leary, 2009). Which emotion is experienced depends on the context and the individual learning history and personality (Elliot & Thrash, 2004; Faustino et al., 2020). Generally, affective states that are experienced as aversive are accompanied by the motivation to avoid them.

The emotions we focus on in the studies presented here are pride and embarrassment as possible markers of an approach to a positive self-image or avoidance of a negative self-image in front of others during task execution. In the clinical study, we examine happiness as an accompanying emotion that is less complex, requires less self-awareness, and reflects the affective state on a more general level (Ekman, 1992; Fredrickson, 2001). In the last study, we assess an affective reaction to social-evaluative stress, including a perceived stress rating in general, as well as embarrassment, anger, and frustration.

Links to self-belief updating

As mentioned above, the predictive processing of incoming information is biased (Mokady & Reggev, 2022). As we approach positive affective states and avoid negative ones, these motivations can bias how we update beliefs in a specific context (Sharot et al., 2023). In turn, observed biases in belief updating may allow conclusions about underlying motivations as they indicate which information was given particular weight. Information about the affective experience during belief updating might further support these indications of the underlying motivation.

Linking observed biases in self-belief updating to self-related motivations (see Table 1.1), self-enhancement should result in a positivity or optimism bias (Mokady & Reggev, 2022). As self-enhancement strives to hold positive self-serving beliefs associated with a positive affect, it goes along with a greater emphasis on positive feedback (Leary, 2007). In the case of positive prior beliefs, self-verification motivation may also be relevant for positively biased belief updating (Mokady & Reggev, 2022). If, for example, I think I am a healthy person and learn that the risk of a certain disease is lower than I thought (positive prediction error), I adjust my expectations particularly strongly, as it confirms the positive global self-belief of being healthy. The confirmation bias in belief updating is especially attributed to the motivation for self-verification as it is linked to the maintenance and reinforcement of existing beliefs (Nickerson, 1998). Self-verification is also discussed in connection with seeking negative self-confirming information in individuals with depression, as confirmation supposedly is rewarding (Mokady & Reggev, 2022). This would result in a negativity bias in self-belief updating. Since the affective experience after negative feedback is usually particularly negative in individuals with depression (Jankowski et al., 2018), the approach motivation is debatable.

To motivate self-improvement, negative feedback is relevant, as it is informative when improvement is needed (Strube, 2012). Whether the relevance of negative feedback also contributes to a stronger incorporation into the self-concept, which would result in a negatively biased updating, remains to be seen. In the case of actual improvement over time, this should be expressed in increasingly positive beliefs about one's abilities, which is, however, not a bias in this case but actually changed abilities.

Avoidance motivation in social situations is associated with negatively biased attention, interpretation, and recollection of social information (Strachman & Gable, 2006). For example, when one holds a self-belief of being inadequate and unimportant and fears rejection, one focuses particularly on social cues that indicate rejection (e.g., a disparaging look) or tends to interpret a neutral look as disparaging. As this kind of threat monitoring increases the likelihood of catching a negative cue, the people who fear rejection the most are also the ones who feel the most rejected and lonely (Strachman & Gable, 2006). Drawing attention primarily to negative social cues suggests that these should be processed preferentially and that self-beliefs tend to be biased in a negative direction (or that existing negative self-beliefs are reinforced). Self-belief updating in contexts that mainly elicit avoidance motivation has barely been studied. Studies reporting a negativity bias (Brotzeller & Gollwitzer, 2024; Ertac, 2011; Zamfir & Dayan, 2022) used

performance contexts to study belief updating about one's ability. Active task execution means there is a possibility to fail, which can lead to social anxiety (Leary, 2007), avoidance motivation (Atkinson, 1957), and threat monitoring (Shechner & Bar-Haim, 2016). This could result in a heightened focus on negative evaluations and negatively biased self-belief updating. Coming from embarrassment research, the previous study from which the current belief updating task was developed showed that participants had increased dwell time of gaze on social-evaluative threat cues when receiving negative feedback, especially with increased embarrassment (Müller-Pinzler et al., 2015). This indicates that the task used here is associated with fear of negative evaluation and threat monitoring. Thus, an underlying motivation to avoid negative social evaluation can be assumed.

To better understand observed biases in self-belief updating, settings in which it is examined will be reviewed more closely first and will be linked to possible underlying motivations that are elicited by these different contexts.

Table 1.1. Overview of self-related motivation and affect associated with self-belief updating

Approach Motivation		Feedback orientation	Affect	Bias
Self-verification	confirm pre-existing self-beliefs	Positive (and negative?)	Consistency, predictability, and security	Confirmation bias
Self-enhancement	achieve positive self-evaluations	Positive	Pride in being a valued person	Positivity bias/ Optimism bias
Self-improvement	develop new skills	Positive and negative (behavior-oriented)	Pride in own accomplishments	?
Avoidance motivation				
Avoid negative evaluation		Negative	Social anxiety, embarrassment, shame	Negativity bias

Note. This is a simplified overview. It does not claim to be exhaustive. Motivations can occur in combination and interact.

1.3.2 Context matters: Self-belief updating under different conditions

Different learning contexts can make certain self-related motivations especially salient, which increases the likelihood of specific emotions being present. This can impact belief-updating behavior in different ways (Wittmann et al., 2016) so that biases in both directions have already been observed. To structure possible influencing factors of the learning context, the belief updating task itself and the task embedding are distinguished.

According to the predictive processing framework, the prior belief and the feedback influence how much a belief is updated (Mokady & Reggev, 2022). If this is translated to learning tasks, the type of self-belief addressed and the type of feedback can be considered possible factors within the task. The **type of belief** addressed can be of rather positive or negative valence on average in a sample. For example, prior self-beliefs about the performance in an empathy test might be relatively positive in a sample of psychology students (this means the type of sample plays a role here, too). More positive prior beliefs (e.g., about one's personality or future) may elicit a stronger motivation to maintain the belief associated with positive affect. This results in positively biased belief-updating (Korn et al., 2012; Sharot et al., 2011).

Furthermore, the type of belief addressed has a certain precision (relatively high or low), making the belief more or less resistant to change in case of conflicting feedback within the task (Lord et al., 1979; Richard & Petty, 2014). The precision of the belief tends to be related to the position of belief within the belief hierarchy from specific to global (Nave et al., 2020). More global beliefs were formed over extended time, so the precision tends to be higher. For example, the global belief "I am a likable person" is usually positively pronounced in a healthy sample and has high precision (Clark, 2013; Hohwy, 2013). Beliefs like "I am good at this task" (newly designed for a study) are specific, and precision is relatively low as the participants have no or only weakly related previous experience to refer back to. If the beliefs addressed have a high precision on average in a sample, such as beliefs about one's IQ in a student sample, belief updating is conservative (Möbius et al., 2011). Also, a bias towards self-enhancement is more pronounced when broad compared to specific traits are addressed (Hughes & Zaki, 2015).

In addition to the self-belief addressed, the **type of feedback** presented to induce a self-belief update is relevant. Central to the belief update's direction is the prediction error valence (which is usually an independent variable controlled for a balanced prediction error presentation). Apart from the valence, the precision of the feedback can be varied with very precise self-related feedback (e.g., "You are better than 61 % of the reference group"; Müller-Pinzler et al., 2015) or vague feedback (e.g., "With a probability of 75%, you are among the top half of performers"; Möbius et al., 2011). Vague feedback offers more scope for individual interpretation and distortion (Jug et al., 2019; Kluger & DeNisi, 1996), which can result in more biased belief updating. Educational psychology differentiates different types of feedback in performance contexts with varying impacts on self-related motivations (Hattie & Timperley, 2007): for example, self-related ("You are …") or behavior-related feedback ("Your answer is …"), social-comparative feedback (e.g., "You are better than 60 % of the group"), feedback in comparison to a criterion (norm-referenced feedback, e.g., "Your IQ is in the average range") or self-referenced feedback (e.g., "You did better than last week."). It has been shown that if social comparison is highlighted, students focus more on their performance compared to others, which also increases the emotional response to success and failure. Conversely, when absolute standards are emphasized, the focus shifts to self-improvement, leading students to pay more attention to their effort and task strategies (Ames & Archer, 1988).

This means that social comparison feedback can activate an approach motivation to be better than others, but it can also create the feeling of failure and thus an avoidance motivation. Thus, different biases can be assumed.

A further distinction concerning the task is the **task engagement**; that is, whether the participants are active or passive during task execution. In the passive case, they receive feedback regarding a self-related domain they cannot actively influence during the task (Sharot et al., 2011). For example, feedback on the probability of getting cancer cannot be altered by effort within the task. In the active case, participants perform a task and receive performance feedback. This means they influence the feedback during the task (or at least believe this in the case of pre-programmed feedback; Brotzeller & Gollwitzer, 2024; Zamfir & Dayan, 2022). Whenever there is an opportunity for individual action, there also exists the potential for failure. Thus, performance contexts can elicit self-related motivations like self-improvement or avoiding failures (Ames & Archer, 1988) along with self-conscious emotions like pride or embarrassment (Lewis, 2008). The fear of failure may be amplified by the **task difficulty**, with high difficulty evoking a sense of being overwhelmed and unsuccessful. Fear of failure is linked to avoidance motivation with a focus on negative feedback (Atkinson, 1957), and a negatively biased belief updating can be assumed.

The contextual embedding of a learning task can further elicit certain self-related motivations and, thus, influence how self-belief updating is biased (Wittmann et al., 2016). A potential factor is whether the task is **socially embedded**, such as having other participants present or making social comparison salient (Müller-Pinzler et al., 2015; Tesser et al., 1988). The mere social presence with potential observation is accompanied by physiological reactions like enhanced skin conductance (Cacioppo et al., 1990). Experimental manipulation of participant's concerns with others' impressions of them or social comparison can elicit the fear of negative social evaluation linked to social anxiety, threat monitoring, and avoidance motivation (DePaulo et al., 1990; Steinmetz et al., 2016; Tesser et al., 1988). Additionally, the presence of an obvious social threat during a task performance, like an audience observing the performance, can further amplify emotional reactions like social anxiety, embarrassment, or shame and the focus on personal errors (Dickerson et al., 2008; Gruenewald et al., 2004). This is accompanied by a physiological stress response with increased cortisol and heart rate (Kirschbaum et al., 1993b). Also, other types of threats increase anxiety and bias self-belief updating (Garrett et al., 2018a). Whereby the time point of the social threat induction, before or during task execution, is relevant. After experiencing failure, individuals tend to self-enhance (Kurman, 2006; Steele, 1988; Tesser & Cornell, 1991). Therefore, a social threat and the perception of failure before the task increases self-enhancement (Hughes & Beer, 2013), which may shift attention towards more positive feedback and influence belief-updating in a self-serving direction. This will be the focus of Study 4. The impact of another observer during task performance on self-belief updating will be addressed in Study 1.

Table 1.2. Overview of possible influences of task and setting characteristics on self-belief updates

Within the task			
Type of self-beliefs addressed			
Valence	Positive		Negative
Precision	High		Low
Position within hierarchy	Global		Specific
Type of feedback			
Valence	Positive PE		Negative PE
Precision	High		Low
Social	Social-comparative		Non-comparative
Task engagement	Active performance		Passive feedback
Task difficulty	High	Low	
Contextual embedding of the task/ experimental manipulation			
Social			Non-social
Presence of others (number? characteristics? role within experiment?) Social comparison (salience? direction?)			
Threat induction			
Yes			No
Type of threat			
Social	Non-social		
…	…		
Time point			
Before task	During task		

Note. This overview does not claim to be exhaustive. It is only intended to provide an initial structure for classifying various biases in self-belief updating. Listing binary characteristics is intended to provide a simple overview and not indicate that the factor is dichotomous. PE = Prediction error.

In summary, the salience of social rewards (e.g., confirmation of positive self-belief) or social threats (e.g., negative evaluation) and the elicited self-related motivation can be seen as a combination of 1.) the self-belief addressed in the task, 2.) the type of feedback, and 3.) the task's embedding (for an overview see Table 1.2). Various motivations and emotions can interact, especially in more complex experimental settings. Depending on the motivations and emotions evoked, self-belief updating can be biased in different ways.

1.4 Self-beliefs in depression and social anxiety

1.4.1 Introduction to self-beliefs in depression and social anxiety

We all hold positive as well as negative self-beliefs. This means negative self-beliefs are not a sign of psychopathology per se. However, in psychopathological states, they tend to manifest on a more global level, making them more prominent, occupying more mental space, and exerting a stronger influence on perception, emotions, and behavior (Barnard & Teasdale, 2014; Beck, 1979). In the following, the two disorders, depression, and social phobia, will be discussed in more detail. Depression and social phobia often occur comorbidly (Adams et al., 2016; Kessler et al., 1999).

In both of these disorders, negative self-beliefs are a key factor in the cluster of symptoms. Depression is one of the most common mental illnesses (Kopala-Sibley & Klein, 2017) and is characterized by affective symptoms like persistent sad, anxious, or "empty" mood, hopelessness, irritability, and feelings of guilt, worthlessness, or helplessness. Physical symptoms are decreased energy, changes in appetite, and sleep disturbances (Kopala-Sibley & Klein, 2017). In addition to concentration and decision-making difficulties, and suicidal thoughts on the cognitive level, negative beliefs about oneself and others are central symptoms of depression (Beck et al., 2024). Negative self-beliefs often revolve around negative self-description, for example, “I am unimportant,” and low self-efficacy, for example, “I cannot do it” (Beck, 1967). The expectation to fail combined with low energy often leads to withdrawal behavior and, therefore, the loss of positive reinforcement (Lewinsohn, 1974). Negative self-beliefs of being helplessly exposed to the environment and having no control, which are generalized to new situations, are important cognitive factors in depressive symptoms (Maier & Seligman, 1976). Metacognitive beliefs regarding the lack of control over one's thoughts are a central element in the metacognitive theory of depression, leading to persistent rumination and, thus, maintenance of symptoms (Wells, 2011). One of the most influential theories in behavioral therapy is Beck's cognitive model of depression (Beck, 1979; Beck et al., 2024). This framework addresses how negative beliefs and patterns of thinking perpetuate depressive symptoms. It includes the cognitive triad, which distinguishes negative beliefs into negative self-beliefs (e.g., "I am ugly"), negative beliefs about the world (e.g., "The world is unfair"), and negative beliefs about the future (e.g., "My symptoms will remain forever"). These beliefs are accompanied by a negatively biased thinking style. Here, various cognitive distortions were identified, such as overgeneralization (drawing broad negative conclusions based on limited evidence) or catastrophizing (expecting the worst possible outcome). It has also been shown that individuals prone to depression show a tendency to focus on, encode, and recall negative information while neglecting positive or neutral information (Dozois & Beck, 2008; Kube, 2023). As beliefs have a learning history, depressive self-beliefs formed in childhood (Dozois & Beck, 2008) are often subject to negative experiences, mostly emotional abuse and neglect, sometimes also sexual or physical abuse (Mandelli et al., 2015). For example, neglect may lead to the self-belief of being unimportant. These deeply ingrained negative self-beliefs may be triggered by challenging life events later on. The concept of emotional schema also includes the

affective component of self-beliefs (Greenberg, 2010). Negative affect is considered central to the learning history, and emotional schemas are formed by connecting highly arousing, traumatic events to the resulting emotional reactions (e.g., shame during humiliation; Greenberg, 2010). When triggered by learned signs associated with the events, it results in the corresponding cognitions (e.g., "I'm worthless") and automatic, rapid, and exceptionally intensive emotional responses (Greenberg, 2010). Therefore, emotional and cognitive reactions in a particular situation may appear excessive to outsiders (e.g., excessive self-deprecation and shame after a minor criticism) but are traceable due to the learning history.

Negative self-beliefs also play a central role in social phobia. They are especially prominent in the (imagined) presence of a social public and, thus, the possible devaluation by others. Individuals with social anxiety usually hold negative self-beliefs, like being socially inept or unlikeable (Leary & Atherton, 1986). These beliefs make them hyper-aware of how others perceive them (Clark & Wells, 1995; Hirsch et al., 2003). The central affect is the fear of being negatively judged by others. This goes along with the anticipation of rejection, criticism, or embarrassment, which heightens anxiety in social situations even more (Clark & Wells, 1995; Heimberg et al., 2014). Threat monitoring in social anxiety includes heightened self-focused attention with monitoring one's performance and one's anxiety symptoms like blushing, as well as monitoring other's evaluation of it (Clark & Wells, 1995). Negative prior self-beliefs and threat monitoring result in a mental representation of oneself from an imagined audience perspective (e.g., blushing, sweating, and clumsy; Heimberg et al., 2014). To counter this representation, people pursue safety behaviors such as avoiding eye contact (e.g., to hide blushing and avoid the expected negative evaluation in the other person's gaze) or avoiding the social situation altogether. It reduces anxiety in the short term; however, it prevents a corrective experience, which perpetuates negative self-beliefs and anxiety in the long term (Wells et al., 1995). Additionally, negative self-beliefs are repeatedly reinforced by post-event rumination, that is, the negatively biased replay of social encounters focusing on one's mistakes (Clark & Wells, 1995).

When examining self-beliefs in social anxiety and depression, certain task characteristics may be more or less likely to trigger negative self-beliefs and the associated affective states. An active performance context could activate self-beliefs of failure and inadequacy. The way a task is socially embedded is especially relevant when studying self-belief in social anxiety. All of this can impact the perceived social threat, feedback processing, and, therefore, the updating of self-beliefs.

1.4.2 Belief updating in depression and social anxiety

Since negative self-beliefs in depression and social anxiety are strongly linked to painful negative emotions and a high overall symptom burden, the question arises about the learning process involved in forming and updating these self-beliefs. More precisely, when beliefs tend to be adapted in response to the environment and when active inference is used to maintain existing beliefs. As mentioned above, a central element of the cognitive model of depression is cognitive distortions. This results in biased information processing that sustains negative self-beliefs (Beck, 1963, 1964, 1979). The theory has been

supported by research showing a more negatively biased updating of beliefs about one's future (Garrett et al., 2014; Korn et al., 2014) or social popularity (Will et al., 2020) in individuals with depression or related symptoms like low self-esteem. In recent years, information processing in mental disorders has been understood as a maladaptive response to prediction errors (Clark et al., 2018; Stephan et al., 2016). In cases of depression, findings have shown diminished updating of self-beliefs in response to positive prediction errors, such as receiving unexpectedly positive feedback in a performance setting or in imagined social interactions (Everaert et al., 2018; Kube et al., 2019). This reduced updating of negative self-beliefs can be subject to maladaptive cognitive mechanisms that devalue disconfirming positive feedback. In line with the concept of active inference, this means re-evaluating the information from the environment to make it fit the prior self-beliefs. Aaron T. Beck described discounting the positive as one of various cognitive distortions that play an important role in cognitive behavioral therapy (Dobson & Dozois, 2021). More recently, this cognitive defense against positive information is described as making oneself cognitively immune to positive information (i.e., cognitive immunization; Kube et al., 2019). It is seen as the central cognitive mechanism for maintaining negative self-beliefs as it makes them so difficult to change. Strategies of cognitive immunization can be, for example, discounting positive feedback by viewing it as an exception (e.g., "Only this one therapist is nice to me; everyone else wants to harm me") or doubting its credibility (e.g., "My therapist is only nice to me because it is her job"; Kube et al., 2019; Rief et al., 2015). As beliefs and affect are linked (Bromberg-Martin & Sharot, 2020; Eldar & Niv, 2015), this recursive influence should be considered part of the belief updating process. Depression is characterized by predominately low mood, which has been linked to less flexibility in adjusting self-beliefs in a positive direction (Karnick et al., 2024; Kube & Korn, 2024). When in a low mood, the corresponding negative learning history is especially present, making positive information seem implausible and negative self-beliefs appear to be the best prediction. Accordingly, withdrawal behavior follows to avoid further disappointment or harm. Addressing emotional meanings and in-session emotional experiences can enhance the effectiveness of psychological interventions to change maladaptive beliefs (Samoilov & Goldfried, 2000). This highlights how affect and cognitive strategies interact when updating maladaptive self-beliefs. Therefore, the affect in association with self-belief updating is given a central role in the work presented here.

In individuals with social anxiety, negative self-beliefs also go along with biased information processing, with the addition that the presence of an (imagined) audience impacts this. In social situations, individuals with social anxiety have more negative expectations and interpret ambiguous social feedback as rather negative, with a tendency to catastrophize (Chen et al., 2020; Smith & Sarason, 1974). Once feedback is given by a computer, arousal and anxiety are less pronounced (Peterburs et al., 2016). For feedback from other people, however, individuals with social anxiety have difficulties in judging fear-inducing information as false and reassuring, disconfirming information as true (Vroling & de Jong, 2009). The negatively biased information processing is also expressed by more accurate

memory of negative information about oneself in individuals with high social anxiety (O'Banion & Arkowitz, 1977). When retrospectively rating their performance in a public speech, highly socially anxious individuals rate themselves more negatively, even when controlling for observable differences in performance and anxiety (Ashbaugh et al., 2005). This means they overestimate how much potential presentation performance deficits are apparent to others. Similar to individuals with depression, negatively-biased information processing prevents maladaptive self-beliefs from being updated, which maintains anxiety (Vroling & de Jong, 2009). Thus, social anxiety is associated with reduced revision of negative beliefs following disconfirming positive information (Everaert et al., 2018). This inflexibility in belief revision is related to dampening positive emotions (Everaert et al., 2020). A positivity bias in non-anxious individuals when learning from social feedback about being liked or disliked is absent in individuals with social anxiety (Button et al., 2015). This reduced learning from positive feedback remains stable, as shown in a one-year follow-up (Koban et al., 2017). When computationally modeling the affective reactivity to performance feedback, individuals with social anxiety show a stronger integration of negative feedback into their feelings about the self (Hopkins et al., 2021; Koban et al., 2017). In summary, negative feedback in social situations has a particularly strong influence on people with social anxiety, and negative self-beliefs are difficult to correct through positive feedback, similar to individuals with depression.

1.5 Studying neurocomputational mechanisms of changing self-beliefs

1.5.1 Computational models of self-related states

To describe the dynamics of self-related states (e.g., self-beliefs, affective states, or state self-esteem) as a function of incoming information, methods that can capture moment-to-moment change are needed. Computational models mimic how the mind processes information, which allows for testing hypotheses about learning, decision-making, and other mental functions. Models are built on theoretical assumptions about how cognitive or affective processes work and express these assumptions as mathematical equations that describe how inputs (e.g., performance feedback) are transformed into outputs (e.g., decisions or actions; Lewandowsky & Farrell, 2011). These equations, in turn, involve parameters that govern how information is processed, and a model is defined by its precise constellation of parameters and equations. Computational models allow the mechanistic interrogation of trial-by-trial variations, which is inaccessible with classical methods based on summary statistics and requires explicitly defining the parameters that drive behavior (Zhang et al., 2020). Varying model parameters and testing which model best describes empirical data (winning model) can provide insight into the mental processes that underpin observed behavior.

One specific field of application is reinforcement learning (Sutton & Barto, 2018), where models capture fluctuations of latent decision variables in learning and decision-making tasks. Here, the expectation is an implicit value given to a certain choice option that is iteratively updated as a function of prediction errors. This model is used to describe choice behavior between usually two options with varying reward schedules of monetary reward (Zhang et al., 2020). An influential computational model was developed by Rescorla and

Wagner (1972), who built on early behaviorist studies of classical conditioning (Pavlov & Anrep, 1927; Watson & Rayner, 1920) to explain how learning happens through prediction errors. This learning process can be represented mathematically, where future expectations (EXP[t+1]) are determined by current expectations (EXP[t]) adjusted by the prediction error (PE[t]). This adjustment is modulated by a learning rate (α). Essentially, the prediction error only reflects the magnitude of the difference between the expected (EXP[t]) and the actual outcome or feedback (FB[t]). The learning rate influences how much this error affects learning, as described by the equation: $EXP[t+1] = EXP[t] + \alpha * PE$, where $PE = FB[t] - EXP[t]$ (Lockwood & Klein-Flügge, 2020; Rescorla & Wagner, 1972). The higher the learning rate, the stronger the weighting of the prediction error for the expectation update.

Computational models such as these can also be applied to the context of self-related learning. For example, one can model the expectation to be liked by another person or the state self-esteem over time as a function of trail-by-trial social feedback (Hopkins et al., 2021; Will et al., 2017). Linking learning parameters with clinical variables allows conclusions about alterations in the self-related learning process with increasing symptoms. For example, learning parameters for negative social-evaluative prediction errors are higher when fear of negative evaluation is high (Hopkins et al., 2021). This suggests a stronger incorporation of negative social evaluation into the learning process. Models can describe binary choices (e.g., "I think you are…": "dull" vs. "witty;" Hopkins et al., 2021) or fluctuations of dimensional variables like ability-beliefs (Zamfir & Dayan, 2022), state self-esteem (Will et al., 2017), or affective states (Charpentier et al., 2016; Rutledge et al., 2014).

When using the Rescorla-Wagner model described above to model self-beliefs as a function of positive and negative social evaluation, model comparison results show that a model with separate learning rates for positive and negative prediction errors can describe the fluctuations better than a model with only one learning rate (Elder et al., 2022; Koban et al., 2017). This suggests that positive and negative prediction errors in social evaluation impact self-beliefs differently. Similarly, a model that includes a bias parameter that captures global beliefs about being liked or disliked biasing the updating process outperforms other models that do not account for biases (Will et al., 2017). When comparing the learning rates between individuals with and without social anxiety in a second step, results show that individuals with social anxiety have higher learning rates for negative prediction errors; this means an overall stronger negativity bias in affective belief updating (Koban et al., 2017). Overall, the learning rate increases with increased depressive-anxiety symptoms, which indicates increased fluctuations of feelings about oneself in response to prediction errors of social evaluation (Will et al., 2017). Aside from adding bias terms or multiple learning rates, the basic Rescorla-Wagner model can be extended by various components to map the complexity of affect or belief updating in social contexts. For example, parameters that map the level of detail or prior knowledge of social groups when updating beliefs about others (Frolichs et al., 2022). This expands the space of different computational mechanisms that can be tested against each other to better understand the underlying process.

In a classic reinforcement learning paradigm, participants typically have no prior beliefs about the reward contingencies of choice options and gradually form an implicit value over time. However, in self-related learning, participants enter the experiment with a learning history and prior beliefs about themselves. Thus, computational models are usually used to describe fluctuations of self-related states rather than the formation of novel beliefs. In the present work, the influence of prior self-beliefs is to be reduced to investigate the formation of relatively novel self-beliefs. Here, in addition to self-belief fluctuations known from previous paradigms, an average shift in one direction over time can also be expected (e.g., from a neutral belief about one's ability to "I am good at this task"). To capture this, different variations of the Rescorla-Wagner model will be considered in the model space, including models that estimate the weight of positive and negative prediction errors separately, as models with learning bias won in previous studies (Koban et al., 2017).

1.5.2 Neural underpinnings of self-belief updating

Predictions, prediction errors, and tracking stimuli's value are the basis of social and non-social learning and belief updating. Neuroscience and computational modeling laid the groundwork for this understanding of predictive processing by formalizing how the brain uses prediction errors as a signal to update its internal models (Dayan et al., 1995; Mumford, 1992; Rao & Ballard, 1999; Rumelhart et al., 1986) and providing first correlates of prediction error processing (Näätänen et al., 1993; Schultz et al., 1997). Today, we know that various brain regions are involved in tracking prediction errors and the value of stimuli.

Reward prediction and prediction errors as the basis of value learning are typically processed in dopaminergic neurons in the ventral striatum, ventral tegmental area, and substantia nigra (Diederen et al., 2016; Ruff & Fehr, 2014; Schultz et al., 1997). Unexpected reward results in increased activity of dopamine neurons from baseline (positive prediction error), unexpected omission of reward results in a decrease in activity (negative prediction error; Ruff & Fehr, 2014; Schultz et al., 1997; Zald et al., 2004). Additionally, functional MRI research has pointed to the ventromedial prefrontal cortex (comprising the medial orbitofrontal cortex and parts of the medial prefrontal cortex) as being involved in computations of a value signal that integrates expected rewards and costs at the time of the decision (Hare et al., 2008; Kable & Glimcher, 2009; Ruff & Fehr, 2014).

Although early findings on prediction error processing and value computation were based on non-social/ non-self-related stimuli, recent functional MRI research suggests that the neural representations of value-related computational processes in both social and non-social contexts share similar basic principles (Ruff & Fehr, 2014). For example, when receiving self-related social feedback, the rewarding component is tracked in the ventral striatum and anterior cingulate cortex (Koban et al., 2023; Korn et al., 2012; Müller-Pinzler et al., 2015; Will et al., 2017). Self-belief updates covary with activity in the ventromedial prefrontal cortex (Kuzmanovic et al., 2016; Will et al., 2017).

As self-belief updating is associated with specific self-related motivations and emotional investment (see section 1.3.1), there are also neural correlates that may reflect these

more self-specific components of belief updating. A recent model (Dixon & Gross, 2021) proposes that self-related processing in the brain is organized across several key networks: The default mode network is responsible for representing self-related content, while a valuation network, which includes the insula, midcingulate cortex, and limbic regions, assesses the emotional value (positive or negative) of these beliefs. Last, the frontoparietal network regulates these self-beliefs and emotional responses in a flexible and context-sensitive manner (Dixon & Gross, 2021). In regard to self-specific representation in value learning, activity in subregions of the anterior cingulate cortex (ACC) depends on whether information about oneself or another agent is processed (Lockwood et al., 2016; Lockwood & Wittmann, 2018).

When updating self-beliefs in the context of approaching positive self-beliefs, a positively biased updating of beliefs about one's future is related to reduced tracking of negative prediction errors in the inferior frontal gyrus (Sharot et al., 2011). As this positivity bias is self-specific, activity in the ventromedial prefrontal cortex is linked to increased positive and decreased negative belief updates, specifically for self-beliefs, and is absent when updating about another person (Kuzmanovic et al., 2016). In addition, the higher the activity in regions including the dorsomedial prefrontal cortex and ventral striatum when confronted with negative prediction errors, the less this undesirable information is integrated, contributing to an overall positivity bias in self-belief updating. When updating beliefs about one's personality in contrast to another's personality, the rewarding component of positive social feedback from peers is related to activity in the ventral striatum and ACC, specifically for self-related feedback (Korn et al., 2012). In this context, the valence-independent discrepancy between expectation and actual feedback from peers is correlated with activity in regions known from the mentalizing network (Frith & Frith, 2003; Mar, 2011), including the medial frontal cortex, inferior frontal gyrus extending into anterior insula, and temporo-parietal junction for both self and other (Korn et al., 2012). The medial prefrontal cortex integrates the tracking of rewards and the valence-independent discrepancy from peers and predicts positively biased self-belief updating (Korn et al., 2012). The ventromedial prefrontal cortex has been discussed as a hub that integrates affective responses with contextual information to guide decision behavior (Roy et al., 2012).

When reporting one's current affective state during learning, state happiness correlates with activity in the anterior insula (Rutledge et al., 2014), which is also associated with interoceptive and emotional awareness (Craig, 2009b; Critchley et al., 2004; Singer et al., 2009). The anterior insula, together with the dorsal mediofrontal cortex and amygdala, is also involved in action monitoring, error processing, and emotional processing (Koban & Pourtois, 2014; Murray, 2007), making it essential for self-belief updating. Especially the anterior insula is discussed to be relevant for the integration of affective states, motivated cognition with outcome information, and social context information (Chang et al., 2013; Koban & Pourtois, 2014; Wager & Barrett, 2017) and thereby essential for more complex situation-specific and conscious emotional states like embarrassment (Koban & Pourtois, 2014). This may be particularly relevant in self-related learning contexts, which will likely

trigger fear of failure and embarrassment. Therefore, the neural processing of aversive feedback and negative prediction errors as a basis for self-belief updating, especially in avoidance contexts, will be looked at next.

Generally, various regions, including the insula, amygdala, medial prefrontal cortex, and midbrain, involved in reward processing were also found in response to aversive stimuli (Leknes & Tracey, 2008; Murray, 2007). Threat cues elicit activation in regions like the amygdala, insula, striatum, and thalamus (Schlund et al., 2010), with the amygdala and posterior insula tracking the magnitude of a potential negative outcome (Canessa et al., 2013). When learning social threats, the learned association between another person and their negative evaluation of oneself is associated with amygdala activity (Davis et al., 2010; Pejic et al., 2013), similar to classical fear condition paradigms (Phelps, 2006). The anterior insula is also associated with error awareness (Ullsperger et al., 2010). Receiving self-related performance feedback in public is associated with stronger recruitment of the mentalizing network, suggesting increased thinking about others' evaluation (Müller-Pinzler et al., 2015). Negative performance feedback, in particular, is associated with an increased coupling of the mentalizing network with the amygdala and anterior insula, as well as increased embarrassment (Müller-Pinzler et al., 2015). Greater activity in the insula, as well as the anterior cingulate cortex, is also associated with greater feelings of social distress in response to social exclusion (Eisenberger et al., 2003; Masten et al., 2009).

When a self-evaluation follows a social-evaluative threat, participants tend to more self-serving evaluations as self-protection, which is related to increased activity in the medial orbitofrontal cortex (Hughes & Beer, 2013). Also, the anterior cingulate cortex is associated with improved access to positive self-evaluation and impaired access to negative self-evaluation following a social-evaluative threat (Hoefler et al., 2015). Positron emission tomography shows coping by increased activation of the endogenous opioid system in the ventral striatum, amygdala, and periaqueductal gray following social rejection feedback (not being liked by others), which was associated with a reduction in negative affect (Hsu et al., 2013).

Neural underpinnings of belief updating in depression and social anxiety

Numerous studies have examined the brain's response to reward prediction errors and learning of reward values in depression, usually with monetary reward stimuli. Consistent with the persistence of negative beliefs despite positive evidence, many of these studies have shown diminished reward learning in depression on the behavioral level (Admon & Pizzagalli, 2015; Kumar et al., 2018; Robinson et al., 2012; Safra et al., 2019). However, some research found no differences in reward learning between individuals with and without depression (Brolsma et al., 2022; Gradin et al., 2011; Rothkirch et al., 2017; Rouhani & Niv, 2019). At the neural level, some studies have reported reduced reward-related prediction error signaling in the ventral striatum (Gradin et al., 2011; Kumar et al., 2008, 2018; Robinson et al., 2012) as well as the ventral tegmental area (Kumar et al., 2018), anterior cingulate cortex, and hippocampus (Chen et al., 2015; Gradin et al., 2011; Kumar et al., 2008). Nevertheless, other studies find no changes in reward prediction error

signaling in depression (Rothkirch et al., 2017; Rutledge et al., 2017). Similarly, for negative or loss-related prediction error processing, some studies show increased activity in the ventral striatum in depression (Ubl et al., 2014), while other studies showed unaltered signaling for negative prediction errors (Kumar et al., 2018; Rothkirch et al., 2017). Thus, there is mixed evidence for altered non-self-specific reward signaling in depression.

With regard to self-beliefs in particular, depression has been linked to heightened tracking of negative prediction errors in the right inferior parietal lobule and inferior frontal gyrus when updating beliefs about one's future (Garrett et al., 2014). In the case of self-threatening feedback, individuals with depression show a stronger sensitivity to social rejection (Kupferberg et al., 2016). When receiving negative feedback of social rejection from peers, activity in the amygdala, anterior cingulate cortex, left anterior insula, and left nucleus accumbens is increased in individuals with depression (He et al., 2020; Jankowski et al., 2018; Kumar et al., 2017; Silk et al., 2014). Heightened neural reactivity to social rejection is associated with more negative self-perceptions, negatively biased information processing (Jankowski et al., 2018), and lower self-esteem (Kumar et al., 2017). In healthy participants, heightened activity in the anterior cingulate cortex during rejection feedback is linked to the development of depressive symptoms in the following year (Masten et al., 2011). Altered insula activity in individuals with depression has been discussed as a neural correlate of more painful emotions (Mutschler et al., 2012) and maladaptive emotion regulation like rumination in depression (Sliz & Hayley, 2012). The induction of depressed mood in healthy participants has been shown to increase reactivity to negative information in the insula (Harlé et al., 2012). Furthermore, endogenous opioid release in the amygdala was reduced in individuals with depression, along with a slower emotional recovery from rejection (Hsu et al., 2015). No difference in brain activity could be shown for social acceptance feedback (Silk et al., 2014).

In individuals with social anxiety, social threat cues elicit increased insula response (Straube et al., 2004, 2005). A meta-analysis of fMRI studies found that individuals with social anxiety exhibit heightened activation in limbic regions, including the amygdala and anterior insula, in response to emotional compared to neutral stimuli (Etkin & Wager, 2007). In particular, negative feedback about one's performance or personal characteristics is associated with increased activity in the medial prefrontal cortex, insula, and amygdala (Blair et al., 2008; Heitmann et al., 2014). When under social-evaluative threat, anxiety levels of individuals with social anxiety correlated with amygdala activity (Tillfors et al., 2001). Activity in the medial prefrontal cortex when receiving performance feedback in public is associated with social anxiety, mediated by the dwell time of gaze on the audience (Müller-Pinzler et al., 2015). This is in line with the fear of negative evaluation in social anxiety going along with an attentional shift in a threat monitoring way (Morrison & Heimberg, 2013). When updating self-beliefs, activity in the insula mediates the effect of negative social feedback on self-belief updates (Koban et al., 2023), contributing to an overall negatively biased learning about the self.

In the present fMRI study with a clinical sample, we focus on the regions of interest within the amygdala and insula, as these are relevant for processing emotional feedback, as well as the midbrain and ventral striatum as a connection to studies that investigated feedback processing in depression in reward contexts.

1.6 Research question

People hold a variety of beliefs about themselves, their characteristics, their values, their competencies and weaknesses, overall, what they base their identity on. These beliefs can be positive or negative, sometimes so negative and global that they are associated with the development of a mental illness. In order to understand these self-beliefs, a look at the learning history is necessary. Usually, we can only understand how a belief arose in retrospect. For example, a patient's biography in psychotherapy has to be explored in retrospect to hypothesize how maladaptive beliefs and associated symptoms originated. When studying the dynamics of self-beliefs in a laboratory setting, participants enter the learning paradigm with prior beliefs about themselves. Depending on the type of self-belief addressed (e.g., global or specific), these prior beliefs are more or less prominent. Many studies examined self-belief updating of rather global self-beliefs, for example, regarding their intelligence or personality (Eil & Rao, 2010; Korn et al., 2012; Sharot et al., 2011). This means that a strong impact of prior beliefs can be assumed. The following studies aim to develop a learning paradigm that captures the process of forming relatively novel self-beliefs. Instead of looking at single unrelated belief updates, the development of beliefs over several interrelated updates was examined through consecutive trials of self-related feedback (Krach et al., 2024). This involved addressing beliefs that are initially vague (i.e., with high uncertainty about the belief), aiming to form beliefs about one's ability over time. For this purpose, a paradigm originally designed for embarrassment induction through public performance feedback (Müller-Pinzler et al., 2015) was developed into a learning paradigm, which can capture the change in current self-beliefs as a function of feedback. In contrast to previous studies in which subjects received self-related feedback while remaining relatively passive, the following studies investigated self-belief formation in a performance context with active task execution. This means there is a (perceived) opportunity to influence the feedback (e.g., by putting in more effort). This active performance context can trigger improvement motivation to gain a sense of competence. It can also lead to fear of failure and a striving to avoid negative social evaluation, as well as increased embarrassment, as shown before (Müller-Pinzler et al., 2015). The task allows to address specific mechanisms of the belief formation process and make potential biases within this process measurable.

In the first study, the learning paradigm was established to test whether it is possible to form novel self-beliefs in an experimental setting and whether this process is subject to distortions. It examined how global prior self-beliefs, specifically self-esteem, as well as individual levels of social anxiety are associated with learning biases and whether a public social context modulates these. In the second study, the neural underpinnings of self-belief formation and its distortions were addressed. In addition, the emotional investment,

which is particularly relevant for self-beliefs, was examined by combining biased learning with self-conscious emotions and pupil dilation as a measure of arousal. The third study also addressed affect and self-belief formation, here in pathological dimensions, by examining a sample with chronic depression with varying comorbid social anxiety. Neural underpinnings of self-belief formation and affect were investigated in this clinical context. In the fourth study, a social public context was applied again. Here, with a focus on a more intense social-evaluative threat that did not occur during learning but was applied in advance. The aim was to investigate how social evaluative stress influences subsequent self-belief formation.

2 Study 1

Negativity-bias in forming beliefs about own abilities

2.1 Abstract

During everyday interactions people constantly receive feedback on their behavior, which shapes their beliefs about themselves. While classic studies in the field of social learning suggest that people have a tendency to learn better from good news (positivity bias) when they perceive little opportunities to immediately improve their own performance, we show updating is biased towards negative information when participants perceive the opportunity to adapt their performance during learning. In three consecutive experiments we applied a computational modeling approach on the subjects' learning behavior and reveal the negativity bias was specific for learning about own compared to others' performances and was modulated by prior beliefs about the self, i.e. stronger negativity bias in individuals lower in self-esteem. Social anxiety affected self-related negativity biases only when individuals were exposed to a judging audience thereby potentially explaining the persistence of negative self-images in socially anxious individuals which commonly surfaces in social settings. Self-related belief formation is therefore surprisingly negatively biased in situations suggesting opportunities to improve and this bias is shaped by trait differences in self-esteem and social anxiety.

---

[1] This study has been published as: Müller-Pinzler, L., **Czekalla, N.**, Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, 9(1), 14416.
My contribution: designing the research, data acquisition, discussion of the data analyses and interpretation of the results, and review and editing of the manuscript.

2.2 Introduction

People examine their own thoughts, behavior and their efficacy, making "corrective adjustments if necessary" (Bandura & Locke, 2003). They develop beliefs about their abilities (Bem, 1965), i.e. the innate and formed capacities that enable them to perform particular tasks successfully, which become strong motivators for subsequent behaviors and are thus fundamental for well-being (Bandura, 2001; Maier & Seligman, 1976; Nolen-Hoeksema, Girgus, & Seligman, 1986; Taylor & Brown, 1988). Already during the formative periods of development, children's beliefs in their academic efficacy, e.g. mathematical or language self-concepts, have the most pervasive direct impact on their judgment of their later occupational efficacy (Bandura, Barbaranelli, Caprara, & Pastorelli, 2001). Not only in childhood, but throughout the entire lifespan self-related beliefs thus shape future performance and behavior (Bandura & Locke, 2003; Kluger & DeNisi, 1996; Krueger & Dickson, 1994). Though intensive research on the influence of peers and societal norms on self-efficacy beliefs has been conducted (Bussey & Bandura, 1999; Eccles, 1989), surprisingly little is known about the learning mechanisms underlying the formation of self-related ability beliefs (Bandura, 2001).

With the present studies we aim to find answers for three central questions: First, how do people process feedback on their abilities and form beliefs about themselves? Second, how do differences in personality impact this process and finally, how does the social context shape such learning?

Previous studies demonstrate that we update our beliefs in response to the feedback we receive. Rather than integrating feedback in a way that results in an accurate representation of the world studies show that self-related information is not perceived objectively (Loewenstein, 2006; Sharot & Garrett, 2016). The perception of self-related feedback is influenced by various motivational factors. Positive beliefs have an intrinsic value (Sharot & Garrett, 2016) as individuals strive to be viewed in a positive and self-serving light (Markus & Wurf, 1987). This culminates in a robust and often replicated positivity bias for learning of self-related information (Eil & Rao, 2010; Kuzmanovic, Jefferson, & Vogeley, 2016; Mobius, Niederle, Niehaus, & Rosenblat, 2013; Sharot & Garrett, 2016; Sharot, Korn, & Dolan, 2011). Particularly, people show increased updates of their self-related beliefs when information was better than expected (positive prediction error) compared to when information was worse than expected (negative prediction error; Sharot et al., 2011). However, all of these studies have focused on self-related belief updating by confronting people with feedback concerning aspects of the self that are often perceived as rather difficult to change (e.g. IQ, likelihood of dying from a disease; Eil & Rao, 2010; Kuzmanovic et al., 2016; Mobius et al., 2013; Sharot et al., 2011). While people might be able to improve in those aspects by long-term training or preventive health strategies they cannot be directly modified by the agent in the course of the experiment. Does this positivity bias thus also apply to the many cases in which the recipient of feedback can immediately alter the behavior that has been appraised? Humans often have the opportunity to improve (Bandura, 2001; Zimmerman, 1990). For example, when processing information about their job or school performance ("Am I good

at my job?"; "Am I a good student?") or sociability ("Am I a likeable person?"), they can directly act to improve them (e.g. by putting more effort in the next task at work or school or acting more prosocially during the next social interaction). There might thus be a difference in how people update their own ability beliefs based on the presence or absence of the perceived opportunity to improve. Situations suggesting little opportunity for improvement may encourage a positivity bias to regulate mood, whilst those suggesting significant opportunities to improve abilities may encourage the processing of negative information to focus effort where it is most needed (Bandura et al., 2001; Jordan & Audia, 2012; Nolen-Hoeksema et al., 1986; Zimmerman, 2002). In order to fully understand how beliefs are formed and updated it is therefore important to explore whether positivity biases also apply in situations suggesting opportunities for change, for instance by providing performance feedback while people develop a novel skill as compared to facing rather unchangeable facts.

An important related question is how people differ in how they form beliefs about their abilities. The functional value of stable self-efficacy beliefs in contrast to "the self-handicapping costs of nagging self-doubts about one's capabilities" has often been discussed (Bandura & Locke, 2003). For example, studies in the field of developmental and educational psychology continuously demonstrate that already at very young age a child's fundamental lack of belief in his/her own ability to achieve – while not lacking in actual abilities – consistently tempers their ambition (Bandura et al., 2001). Self-related beliefs have the potential to imbue perception and interpretation of feedback (e.g. confirming prior beliefs; Blascovich & McFarlin, 1981; Swann, 1983) and thereby impact consecutive behavior (i.e. task persistence and effort; Kluger & DeNisi, 1996; Krueger & Dickson, 1994; Shrauger & Rosenberg, 1970). The impact of negative self-related beliefs might be even more detrimental in individuals with mental health conditions like depression (Moore & Fresco, 2012) and social anxiety disorder (Garner, Mogg, & Bradley, 2006; Hirsch & Mathews, 2000; Vroling & De Jong, 2009). In such clinical conditions negative beliefs can lead to reduced intrinsic motivation or avoidance behavior and thus exacerbate a self-perpetuating cycle of negative self-related thoughts (Goldin, Manber-Ball, Werner, Heimberg, & Gross, 2009; Heimberg, Brozovich, & Rapee, 2010; Leary & Atherton, 1986). It is therefore important to consider interindividual differences in personality to unravel potential maladaptive learning biases and mechanisms specific for self-related beliefs.

The social context itself plays another crucial role with respect to the formation of self-related beliefs. Being in public changes how people perceive and evaluate their own behavior (Müller-Pinzler et al., 2015; Steinmetz et al., 2016) and it is argued that the presence of other individuals increases arousal, implicating behavioral consequences (Triplett, 1898; Zajonc & Sales, 1966). Humans do not only differ in their general self-related beliefs but also in their specific beliefs to be capable of coping with public situations. Especially socially anxious individuals fear social evaluation and feel unable to make the desired impression in a social context (Leary, 2007; Leary & Kowalski, 1995; Morrison & Heimberg, 2013). Thus, the social context elicits negative cognitions and

emotions that are thought to shape self-efficacy beliefs (Bandura & Locke, 2003). Our aim is thus to examine how prior beliefs about the self impact how individuals learn about their own abilities in a performance situation and how the social context in which individuals perform and receive feedback, e.g. feedback provided under observation or in privacy, shapes self-related learning.

Introducing the "Learning of own performance" (LOOP) task (see Figure 2.1), we examined in three studies how people update self-related beliefs in an ability domain that is novel for them, i.e. cognitive estimation (such as estimating the weights of animals), unlike e.g. mathematical skills for which people hold strong and rigid prior beliefs about their potential capabilities. We implemented a performance-feedback-loop that mimics everyday life performance situations. Inferring prediction error (PE) learning rates by fitting computational learning models we assessed the modulatory influence of self-relatedness, prior beliefs, and the social context on belief updating. Our hypotheses were that when learning about the self, the weight of self-related negative feedback would be increased, because negative feedback gains specific importance for behavior regulation by signaling a demand to increase task-related effort. This led us to predict that this effect would be absent for non-self-related feedback. Second, we assessed whether prior beliefs about the self modulate self-related belief-formation. Here, our expectation was that self-esteem and social anxiety would shift updating behavior in line with a confirmation bias. As suggested by prior studies this implies that individuals higher in social anxiety would show increased biases towards negative information (Button et al., 2015; Koban et al., 2017). Third, we expected the negativity bias in social anxiety to be augmented by a social context, i.e. the presence of an evaluative audience, which triggers social fear related cognition and behavior.

2.3 Results

Experimental Design. *Experiment 1: Agent-LOOP.* The LOOP task formed the main frame for three separate experiments. In experiment 1 we implemented the LOOP task manipulating the "Agent" of the estimation performance to assess how participants learned about themselves (Self condition) compared to learning about another (Other condition). In doing so we aimed to provide an answer for our first main question: how do people process feedback on their abilities and form beliefs about themselves (as compared to learning about another person as a control condition)? Participants were invited in pairs to a study on cognitive estimation. The estimation tasks involved answering estimation questions while receiving manipulated relative performance feedback for each question. Participants took turns in performing the task themselves or allegedly observing the other person performing, while continuously indicating the expected performance (EXP ratings) for the upcoming trial in a High Ability condition and a Low Ability condition (resulting in four feedback conditions: Agent condition (Self vs Other) × Ability condition (High Ability vs Low Ability); see Methods for a detailed description of the task). Our second question, how differences in personality impact self-related learning, was

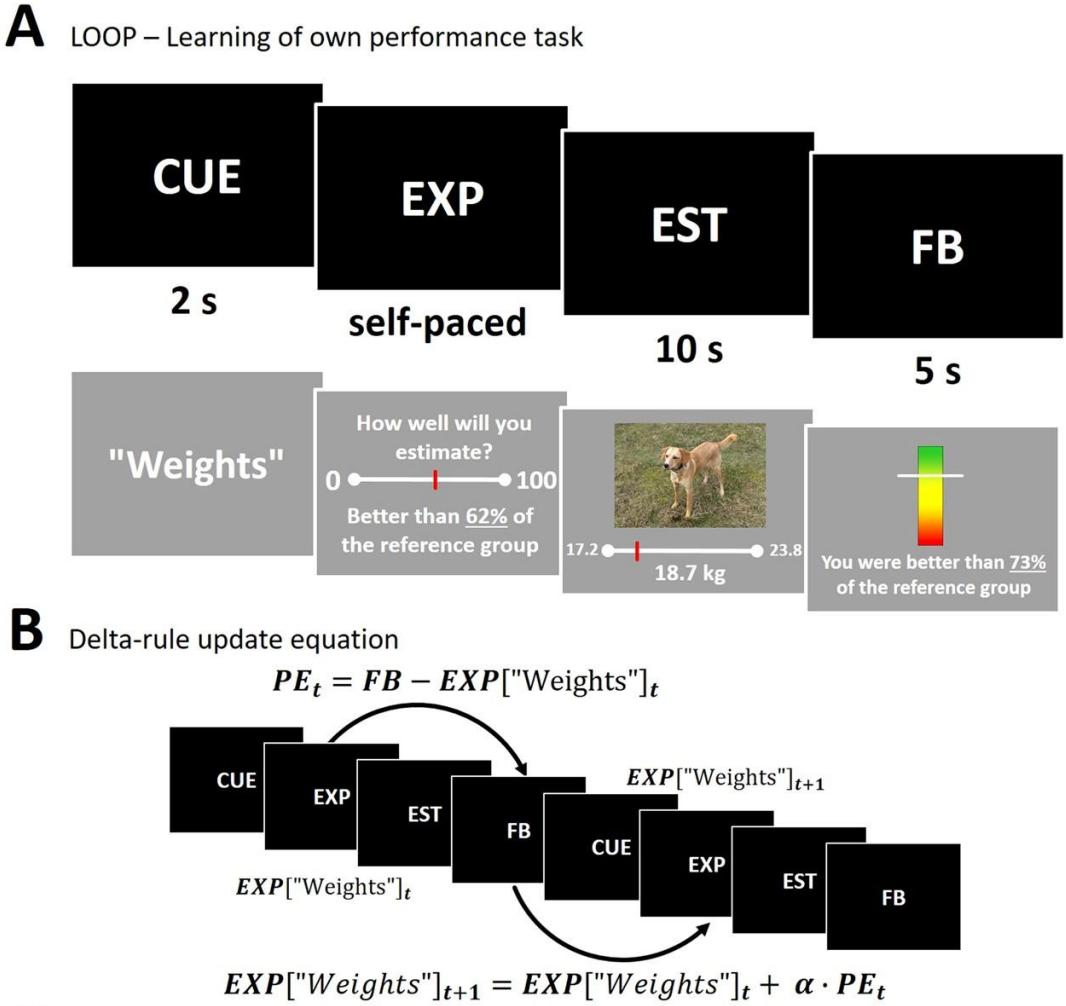


Figure 2.1. Trial sequence, modeling of learning behavior, and experimental factors of the experiments. **(A)** A cue (CUE) in the beginning of each trial indicated the following estimation category. After providing their performance expectation ratings (EXP) participants received an estimation question (EST), followed by the corresponding performance feedback (FB). **(B)** EXP ratings were modeled by means of Rescorla-Wagner delta-rule update equations with different learning rates (α, see 2.5 Methods) taking into account trial-by-trial prediction errors (PE_t) in response to the provided FB. **(C)** In three experiments we assessed the impact of two experimental factors. The "Agent" was manipulated within subjects in the Agent-LOOP task in experiments 1 and 3 and the "Audience" was manipulated in a between-subject design in the Audience-LOOP task (experiment 2) as well as between the Private and the Public group of the Agent-LOOP task (experiment 1 vs experiment 3).

investigated across all three experiments. In experiments 1 we assessed a person's general sense of self-competence or self-esteem via the Self-Description Questionnaire before participants were involved in the estimation task (SDQ-III; Marsh & O'Neill, 1984). We also assessed social interaction anxiety via the Social Interaction Anxiety Scale (SIAS; Mattick & Clarke, 1998). For more details on the sample's questionnaire data see Supplementary Table 1.1.

Experiment 2: Audience-LOOP. In experiment 2 we implemented another version of the LOOP task, to answer our third main question: how does the social context shape self-related learning? We now assessed the impact of the presence of an audience, i.e. being in public or not, on self-related learning in a between-subject design (Figure 2.1C). Participants were invited alone and were randomly assigned to one of two experimental groups (Private vs Public group; see Methods section for further details) resulting in four experimental conditions (Ability condition (High Ability vs Low Ability) × Audience group (Private vs Public)). Here again social interaction anxiety scores served to assess how differences in personality impact self-related learning and specifically how this is modulated by the social context.

Experiment 3: replication and extension. We conducted a third experiment again implementing the Agent-LOOP task (experiment 1), while introducing publicity in a more minimal fashion compared to the Audience-LOOP (experiment 2). With this task variant we aimed to replicate the previous findings as well as to provide evidence for the specificity of the audience effect for self-related learning compared to learning about another person. Self-esteem and social interaction anxiety scores were assessed as described above.

Model-free Behavioral Analysis. We first performed a model free analysis to capture the basic effects we see in our behavioral data (see 2.5 Methods section for further details). For the Agent-LOOP in experiment 1 the Trial × Ability condition × Agent condition ANOVA revealed a significant main effect of Ability condition ($F_{(1,22)}$ = 215.26, p < .001) and interaction of Trial × Ability condition ($F_{(24,528)}$ = 31.43, p < .001) reflecting that participants adapted their EXP ratings over time according to the feedback provided in each Ability condition (see Figure 2.2). The significant main effect of Agent condition ($F_{(1,22)}$ = 15.24, p = .001) and interaction of Agent condition × Ability condition ($F_{(1,22)}$ = 4.65, p = .042) both indicate that participants evaluated their own performance more negatively than the other's performance, specifically in the Low Ability condition. There was no significant interaction of Trial × Agent condition × Ability condition ($F_{(24,528)}$ = 0.99, p = .476).
For the Audience-LOOP (experiment 2) the Trial x Ability condition x Audience ANOVA revealed a significant main effect of Ability condition ($F_{(1,57)}$ = 261.56, p < .001) and interaction of Trial × Ability condition ($F_{(29,1653)}$ = 39.84, p < .001) indicating that participants adapted their EXP ratings over time, while there was no significant impact of the Audience on EXP ratings (main effect of Audience: $F_{(1,57)}$ = 0.09, p = .767; Audience × Ability condition: $F_{(1,57)}$ = 0.15, p = .700; Audience × Ability condition × Trial: $F_{(29,1653)}$ = 1.00, p = .467).

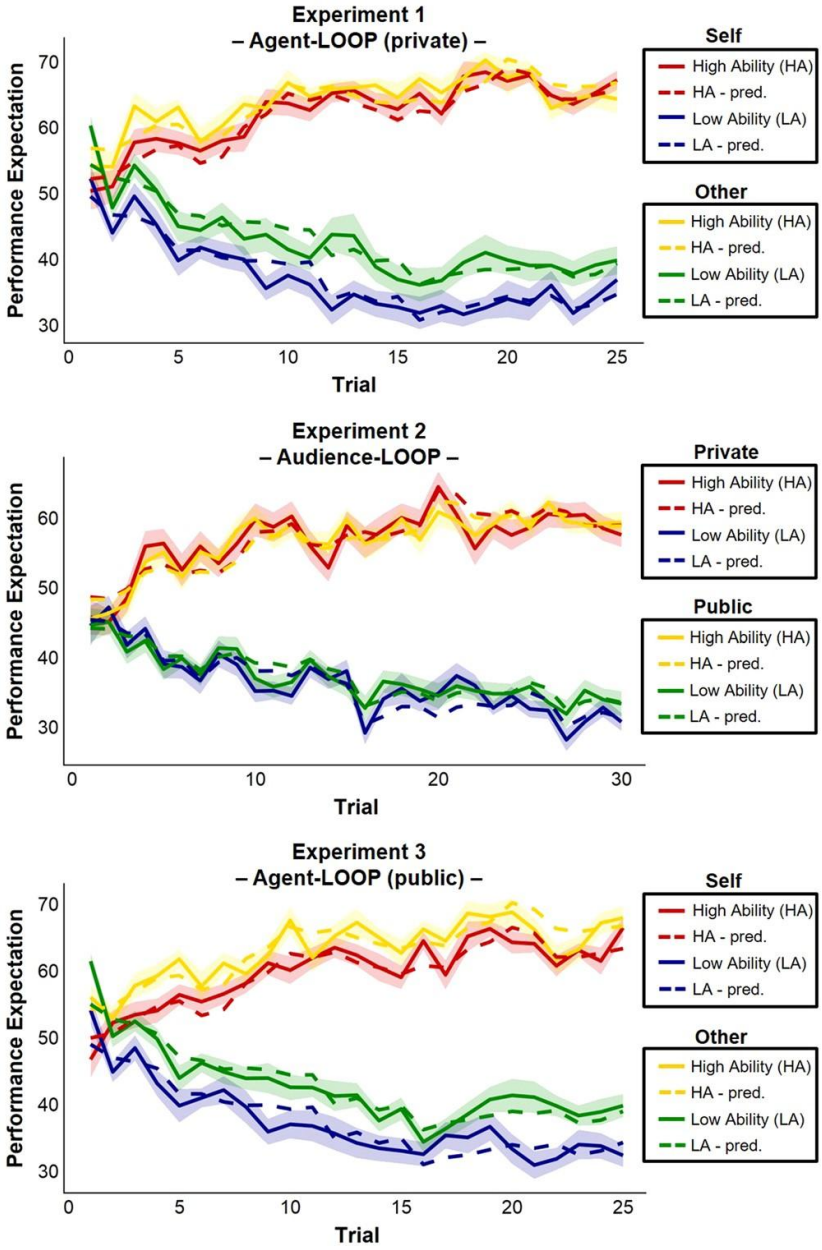


Figure 2.2. Predicted and actual performance expectation ratings across time. The behavioral data of the three experiments (averaged across subjects) indicate that participants adapted their performance expectation ratings (solid lines) to the provided feedback, thus learning about their allegedly distinct performance levels in the two ability conditions. In the Agent-LOOP (top and bottom) participants evaluated their own performance more negatively than the other's performance. Our valence specific learning model captured the participants' behavior for all experiments. Shaded areas represent the standard errors for the actual performance expectations for each trial. Predicted data (pred.) are represented by the dashed lines.

For the public version of the Agent-LOOP in experiment 3, we replicated the findings of experiment 1 (main effect Ability condition: $F_{(1,28)} = 182.99$, $p < .001$; interaction of Trial × Ability condition: $F_{(24,672)} = 36.80$, $p < .001$). Similarly, there was a significant main effect of Agent condition ($F_{(1,28)} = 18.49$, $p < .001$), while the interaction of Agent condition × Ability condition ($F_{(1,28)} = 2.18$, $p = .151$) and the Trial × Agent condition × Ability condition interaction ($F_{(24,672)} = 1.12$, $p = .316$) failed to reach significance, indicating that participants evaluated their own performance more negatively than the other's

performance independently of the ability condition. The combined analysis of the public and private Agent-LOOP (experiment 1 and 3) confirmed the results of the Audience-LOOP by showing that Audience did not have any significant effects also with regards to the additional Agent condition (all *ps* > .439). The remaining effects stayed consistent with the separate analyses of experiment 1 (for more details see 2.7 Supplementary Results).

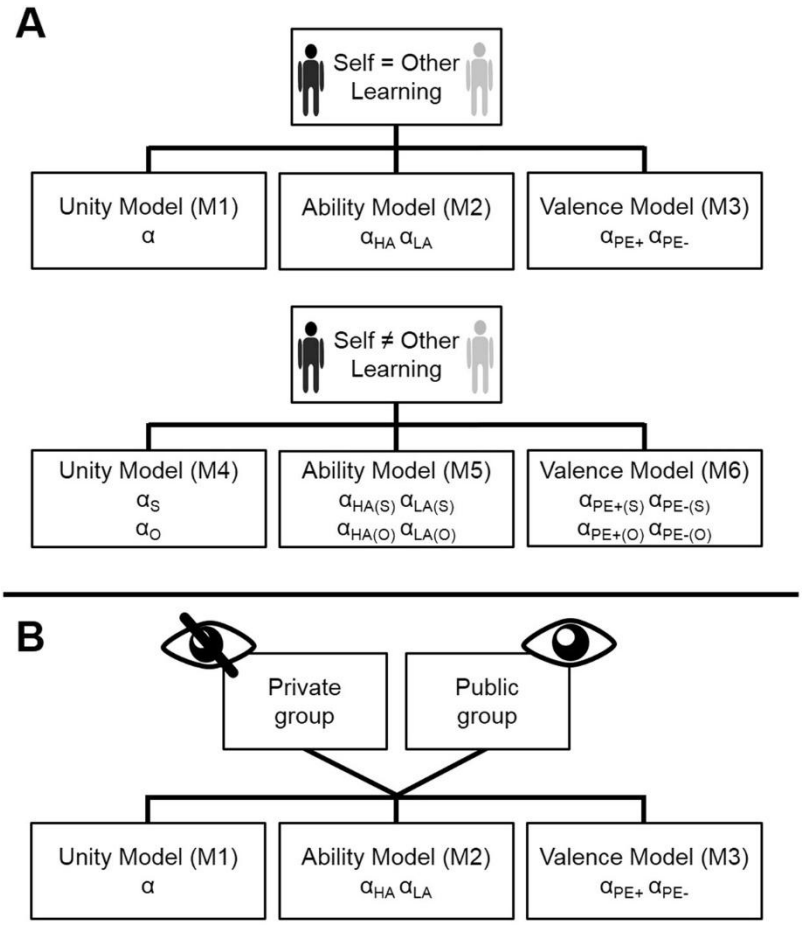


Figure 2.3. Structure of the model space for the three experiments. **(A)** In the Agent-LOOP task (experiments 1 and 3) we distinguished two factors impacting learning rates: the agent (Self vs Other) and the impact (no impact: Unity Model) of the ability condition (Ability Model) or valence (Valence Model). **(B)** In the AudienceLOOP task the impact of the ability condition or valence on learning rates was assessed within the Private and the Public group separately. For a more detailed description of the model space including initial values for the performance expectations see 0 Supplementary Methods.

Model Selection for Computational Models of Learning Behavior. To see whether a learning model can capture the participants' behavior and allows us to summarize the data using principled parameters such as learning rates, we performed a model comparison (see Figure 2.3). Our model space contained three main models varying with regards to their assumptions about biased updating behavior when learning about the self (see Figure 2.3). The simplest learning model used one single learning rate for the whole behavioral time course for each participant, thus not assuming any learning biases [$EXP_{t+1} = EXP_t + \alpha_{Uni}\ PE_t$, while $PE_t = FB_t - EXP_t$; Unity Model]. The second model, the Ability Model, contained a separate learning rate for each of the ability conditions, assuming that participants would show different updating behavior in the High Ability condition (α_{HA}) vs Low Ability condition (α_{LA}). The third model, the Valence Model, included separate learning

rates for positive PEs (α_{PE+}) vs negative PEs (α_{PE-}) across both ability conditions, thus suggesting that the valence (positive vs negative) of the PE biases self-related learning rather than the ability condition itself. In the Agent-LOOP task (experiments 1 and 3) the distinction between learning about oneself vs another person was introduced as a second factor in the model space resulting in three additional models. Model 4 corresponded to the Unity Model with separate learning rates for the self ($\alpha_{Uni(S)}$) and the other person ($\alpha_{Uni(O)}$). Model 5 was the extension of the Ability Model distinguishing between learning about the self ($\alpha_{HA(S)}$, $\alpha_{LA(S)}$) and the other person ($\alpha_{HA(O)}$, $\alpha_{LA(O)}$), resulting in four different learning rates. Model 6 extended the Valence Model by separate learning rates for oneself ($\alpha_{PE+(S)}$, $\alpha_{PE-(S)}$) and the other person ($\alpha_{PE+(O)}$, $\alpha_{PE-(O)}$). To test if the participants' EXP ratings could be better explained in terms of prediction error learning as compared to stable assumptions in each Ability condition, we included a simple Mean Model with a mean value for each task condition (two values for the Audience-LOOP (Model 4) and four values for the Agent-LOOP (Model 7)).

For the Agent-LOOP – implementing model comparison across experiment 1 and 3 – the Valence Model with separate learning rates for Self vs Other (Model 6) received the highest sum PSIS-LOO score out of all models (for all PSIS-LOO scores see Table 2.1, Supplementary Tables 2.2 - 2.3; for a more detailed description of the model space see 2.7 Supplementary Methods). BMS resulted in a protected exceedance probability of $pxp = .998$ (excluding flawed PSIS-LOOs: $pxp_{LOOcorr} > .999$) for Model 6 and a $BOR < .001$ (excluding flawed PSIS-LOOs: $BOR_{LOOcorr} < .001$).

For the Audience-LOOP (experiment 2), there was a clear indication that the Valence Model (Model 3) outperformed all other models according to BMS. Across the Private and Public groups, the protected exceedance probability for the Valence Model was $pxp > .999$ ($pxp_{LOOcorr} > .999$). The BOR was $BOR < .001$ ($BOR_{LOOcorr} < .001$).

Taking into account that model comparisons consistently favored the Valence Model across experiments (Model 6 for the Agent-LOOP and Model 3 for the Audience-LOOP) the Valence Model was selected for all further analyses of learning parameters. Model selection thus revealed that a learning model far surpasses a mean model without learning, and that amongst the learning models, those assuming different learning rates for positive and negative PEs performed best, confirming that it is important to distinguish how positive and negative information is processed. This allowed us to specifically test our main hypotheses of difference in learning about the self vs the other with respect to negative in contrast to positive PEs.

The time courses of EXP ratings predicted by our winning model successfully captured trial-by-trial changes in EXP due to PE updates within each of the ability conditions at the individual subject level ($R^2 = 0.37 \pm 0.24$; $M \pm SD$) supporting the validity of the model in describing the subjects' learning behavior. Posterior predictive checks also confirmed that the winning model captured the core effects in our model free analysis by showing that behavioral analysis on the predictions recapitulates the tendency towards more negative performance expectations for the other that was core to our data (see 2.7 Supplementary Results and Figure 2.2).

Model	PSIS-LOO	LOO-SE	LOO-Diff (SE-Diff)	% of $k\hat{}$ > 0.7	No. Est. Parameters
Agent-LOOP (Experiments 1 and 3)					
Self = Other					
Unity Model (M1)	−2380.1	247.8	135.4 (63.7)	0.1	5
Ability Model (M2)	−2336.5	261.5	91.7 (42.4)	0.3	6
Valence Model (M3)	−2320.5	259.0	75.7 (49.4)	0.2	6
Self ≠ Other					
Unity Model (M4)	−2376.2	254.8	131.5 (54.6)	0.4	6
Ability Model (M5)	−2330.7	263.3	85.9 (42.8)	1.2	8
Valence Model (M6)	−2244.8	283.5	—	0.3	8
Mean Model (M7)	−2953.6	190.3	708.9 (123.3)	0.0	4
Audience-LOOP (Experiment 2)					
Unity Model (M1)	−708.2	145.1	213.1 (35.8)	0.1	3
Ability Model (M2)	−570.2	150.0	75.0 (26.8)	0.3	4
Valence Model (M3)	−495.2	150.9	—	0.1	4
Mean Model (M4)	−1189.5	124.9	694.4 (61.3)	0.0	2

Table 2.1. Model comparisons. *Note*. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (Valence Model) and standard errors of differences; percentage of $k\hat{}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al., 2017); No. Est. Parameters = number of estimated parameters in the model.

Learning Parameters. *Experiment 1: Agent-LOOP.* Participants showed higher learning rates when learning about themselves compared to learning about another person (main effect of Agent: $F_{(1,22)} = 5.23$, $p = .032$). There was no main effect of PE Valence ($F_{(1,22)} = 0.90$, $p = .354$), but the significant interaction of Agent × PE Valence ($F_{(1,22)} = 5.49$, $p = .029$) suggested that there was a bias of updating towards negative information when learning about the self ($t_{(22)} = 1.79$, $p = .088$, $M(\alpha_{PE-(S)}) = 0.14$, $SD = 0.09$; $M(\alpha_{PE+(S)}) = 0.12$, $SD = 0.06$). Learning about another person's performance did not show a significant bias towards negative valence ($t_{(22)} = -0.71$, $p = .484$; $M(\alpha_{PE-(O)}) = 0.10$, $SD = 0.07$; $M(\alpha_{PE+(O)}) = 0.11$, $SD = 0.08$; see Figure 2.4).

Experiment 2: Audience-LOOP. The results of the Audience-LOOP replicated the updating bias towards negative self-related information (main effect of PE Valence: $F_{(1,57)} = 12.64$, $p = .001$; Private: $M(\alpha_{PE-}) = 0.10$, $SD = 0.08$, $M(\alpha_{PE+}) = 0.07$, $SD = 0.06$; Public: $M(\alpha_{PE-}) = 0.08$, $SD = 0.08$, $M(\alpha_{PE+}) = 0.06$, $SD = 0.06$). We, however, did not find any differences in learning rates between the Private and the Public group ($F_{(1,57)} = 1.14$, $p = .290$), nor a significant interaction of Audience × PE Valence ($F_{(1,57)} = 0.35$, $p = .559$), suggesting that being in public might not affect the level of updating in response to negative or positive information per se.

Experiment 3: replication and extension. We again found a significant interaction of Agent × PE Valence ($F_{(1,28)} = 15.45$, $p = .001$), replicating our previous findings of a bias towards negative information, when learning about the self ($t_{(28)} = 3.57$, $p = .001$; $M(\alpha_{PE-(S)}) = 0.15$, $SD = 0.10$; $M(\alpha_{PE+(S)}) = 0.09$, $SD = 0.05$) and no bias towards negative valence when

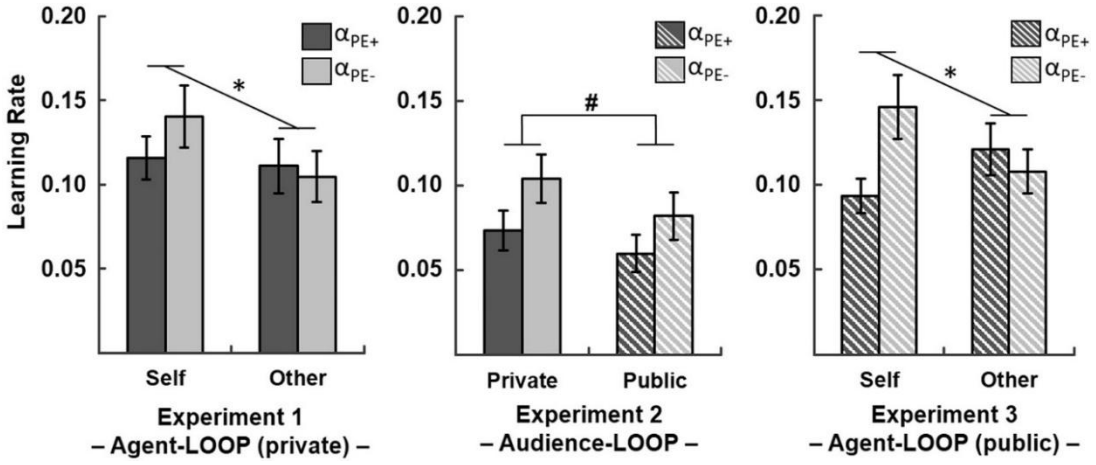


Figure 2.4. Learning rates across the three experiments. The learning rates derived from the Valence Model indicate that there was a bias towards increased updating in response to negative prediction errors (α_{PE-}) in contrast to positive prediction errors (α_{PE+}) across all three experiments. This effect was only present when learning about the self (see left and right) and independent of the social context. Bars represent mean learning rates, error bars depict +/− 1 standard error; * indicates a significant interaction effect of PE Valence × Agent; # indicates a significant main effect of PE Valence across Audience groups.

learning about the other person ($t_{(28)}$ = −1.35, p = .132; $M(\alpha_{PE-(O)})$ = 0.11, SD = 0.07; $M(\alpha_{PE+(O)})$ = 0.12, SD = 0.08). Unlike in experiment 1 learning rates did not differ between the Self and the Other condition ($F_{(1,28)}$ = 0.13, p = .718), due to slightly increased learning rates for the Other condition in the Public group. The main effect of PE Valence reached significance ($F_{(1,28)}$ = 5.25, p = .030), but was driven by a strong bias towards negative valence only in the Self condition. Considering the estimated learning rates of the private Agent-LOOP in experiment 1 and the public version in experiment 3 for the assessment of audience effects, the main effect of PE Valence ($F_{(1,50)}$ = 4.99, p = .030) as well as the interaction of Agent × PE Valence ($F_{(1,50)}$ = 19.01, p < 0.001) remained significant, while the main effect of Agent still failed to reach significance ($F_{(1,50)}$ = 1.98, p = .166). Interestingly, replicating the results of the Audience-LOOP we could not find a main effect of Audience (Audience: $F_{(1,50)}$ < 0.01, p = .966) or any interaction effects (Audience × Agent: $F_{(1,50)}$ = 0.67, p = .416; Audience × PE Valence: $F_{(1,50)}$ = 0.68, p = .414; threefold-interaction Audience × PE Valence × Agent: $F_{(1,50)}$ = 2.44, p = .125). This again suggests that the presence of an audience might not affect updating in response to negative or positive information per se.

Finally, cumulative Bayesian analysis suggests that across all three experiments there was extremely high evidence (Jeffreys, 1961) for a negative valence bias (Bayes $Factor_{10}$ = 19081.7; effect size δ = -0.68, 95%-confidence interval (CI) = [-0.41/-0.95]). Even when adopting an informed prior in favor of a positivity bias (medium mean effect size = 0.5; standard deviation = 0.25), as has been suggested by various studies (Sharot & Garrett, 2016), there still was very strong support for a negativity bias in our data (Bayes $Factor_{10}$ = 94.4; effect size δ = -0.36, CI = [-0.18/ -0.55]).

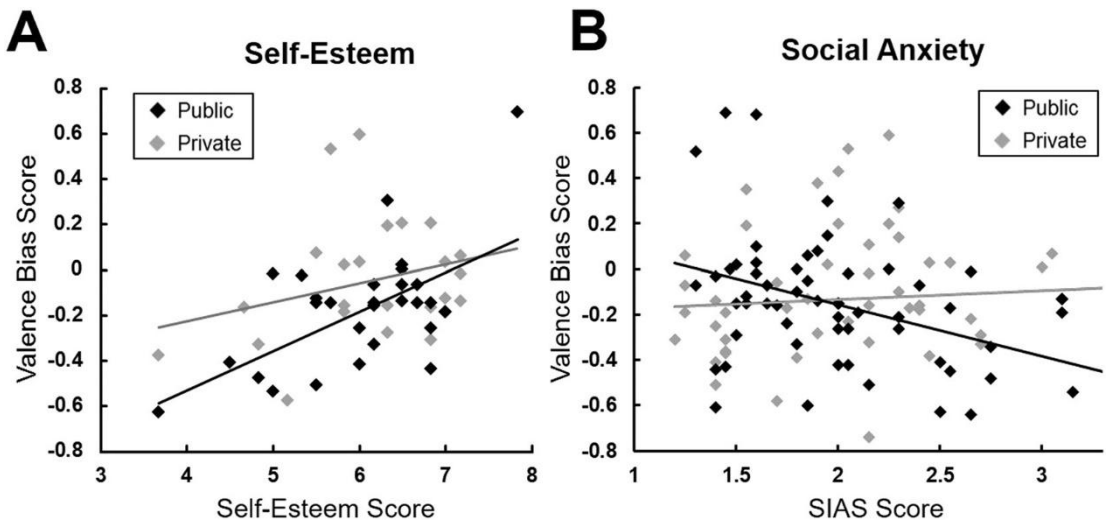


Figure 2.5. Correlation plots of self-related Valence Bias Scores and social anxiety as well as self-esteem for the public and private groups. **(A)** Increased trait self-esteem (SDQ-III score) was associated with a decrease in the negative updating bias about the self in the Public (experiment 3) and the Private group (experiment 1). **(B)** Trait social anxiety (SIAS score) was associated with increased self-related learning biases towards negative information in the Public groups but not the Private groups (across all experiments).

Associations of Learning Behavior with Self-Esteem and Social Anxiety. Partial correlations of Valence Bias Scores (Valence Bias Score = $(\alpha_{PE+(S)} - \alpha_{PE-(S)})/(\alpha_{PE+(S)} + \alpha_{PE-(S)})$; similarly for other-related learning; Niv, Edlund, Dayan, & O'Doherty, 2012; Palminteri, Lefebvre, Kilford, & Blakemore, 2017) and EXP ratings indicated that Valence Bias Scores successfully captured behavioral variance between individuals for all three experiments: Agent-LOOP (experiment 1): r_{part} = .71, p < .001, Audience-LOOP: Private: r_{part} = .78, p < .001, Public: r_{part} = .86, p < .001, Agent-LOOP (experiment 3): r_{part} = .47, p = .006. Thus, individuals with more negative Valence Bias Score ended up with lower self-related performance expectation in the end of the task (controlled for the initial expectations).

The valence bias in self-related learning we found across all three experiments (Valence Bias Score) was negatively associated with interindividual differences in self-esteem in the Agent-LOOP task, $r_{(52)}$ = .44, p = .001 (across experiment 1 and 3; see Figure 2.5A). This indicates that individuals with lower self-esteem showed a stronger valence bias in learning from negative PEs compared to positive PEs. Bayesian analysis corroborated this finding and showed strong evidence for an association of self-esteem and Valence Bias Score (Bayes $Factor_{10}$ = 30.0; effect size δ = 0.44, CI = [0.18/ 0.63]) but the data was inconclusive with regard to a modulating effect of Audience (Bayes $Factor_{10}$ = 0.6).

When assessing the impact of trait social anxiety on updating in response to negative vs positive PEs, we found that the Valence Bias Score was significantly negatively associated with SIAS scores in the Public groups ($r_{(58)}$ = -0.39, p = .002; across experiment 2 and 3), while there was no association in the Private groups ($r_{(50)}$ = 0.06, p = .669; across experiment 1 and 2; difference of correlations: z = 2.39, p = .018). This indicates that individuals higher in social interaction anxiety shifted their updating behavior more strongly towards learning from negative information, specifically when they were in a public

context. This association of the SIAS with learning biases in the public but not in the private context was not present for learning rates for the other person's performance (Private: $r_{(23)}$ = -0.03, p = .897; Public $r_{(29)}$ = -0.21, p = .263; difference in correlations: z = 0.64, p = .524). Bayesian analyses revealed moderate support that the association of SIAS scores and self-related Valence Bias Scores was modulated by the Audience (Bayes $Factor_{10}$ = 7.2). We could find strong evidence that SIAS scores were negatively associated with the Valence Bias in the Public group (Bayes $Factor_{10}$ = 14.4; effect size δ = -0.39, CI = [-0.14/ -0.58]), while there was support for the absence of that effect in the Private group (Bayes $Factor_{10}$ = 0.2; effect size δ = 0.06, CI = [-0.21/ 0.32]).

2.4 Discussion

In a series of three consecutive experiments we explored how individuals update beliefs about their own abilities, and contrasted this against how they update beliefs about others. We aimed to disentangle situational, motivational, and interindividual factors to better understand the nature of learning biases and their relevance for the development of self-concepts. With regard to our first main question, we found that individuals show an updating bias towards negative information about their own performances. When people witnessed feedback about the performance of others, they were as sensitive to positive as negative information. When finding out about their own performance they learned most from negative feedback, which updated their performance estimates more than the positive feedback.

The self-related negativity bias in our findings stands in opposition to the view that self-related learning is positively biased in general (Sharot & Garrett, 2016) and even when adopting a biased prior in favor of a positivity bias, our study provides clear evidence for a negativity-bias in the LOOP task. Hence, one could argue that these results should motivate a closer look on the specific features of tasks applied to examine biases in self-related learning. We argue that several features of the LOOP task introduced here differentiate our task from those that provide evidence in favor of a general positivity bias (Eil & Rao, 2010; Korn, Prehn, Park, Walter, & Heekeren, 2012; Mobius et al., 2013; Sharot et al., 2011).

The specificity of findings on self-related learning suggests that when individuals learn about their own abilities, in contrast to learning about another person, unique motivational factors come into play and shape the way of thinking about and learning from self-related feedback. Subjective desirability of information is considered a constituting factor leading to a positivity bias (Sharot & Garrett, 2016). Such biased updating is in line with the common phenomenon of overconfidence or the so-called "better-than-average" effect, describing the phenomenon that people tend to judge their own performance as better than the average performance (Brown, 1986, 2012). While individuals are typically inclined to hold a positive view of themselves (Sedikides & Gregg, 2008), which can shape the abovementioned situational desirability of positive information, the nature of the performance situation determines how individuals achieve a positive view of the self in the long run. Depending on the situation, two distinct and almost oppositional motives could

in principle color how we process feedback: self-enhancement (i.e. the tendency to evaluate the self positively by either augmenting the positivity or diminishing the negativity of the self-concept) and self-improvement (i.e. the tendency to improve one's own performance to maintain a positive self-evaluation; Jordan & Audia, 2012; Sedikides & Gregg, 2008).

Studies typically find a positivity bias in updating in line with self-enhancement motives when individuals are confronted with feedback in a personally relevant domain (e.g. IQ, health status), but the content (e.g. being intelligent or not) is rather unchangeable during the experiment (Eil & Rao, 2010; Korn et al., 2012; Kuzmanovic et al., 2016; Sharot et al., 2011). Here, self-enhancement motives are triggered because a negative self-related belief in the personally relevant domain would pose a threat for the individual (Jordan & Audia, 2012; Sedikides & Gregg, 2008). Being unable to change the actual outcome with regards to the self, self-enhancement remains the only behavioral option to fulfill the wish for a positive self-view (Jordan & Audia, 2012) and thus might increase the motivation for a positive updating bias when confronted with new information about the self (Eil & Rao, 2010; Swann, 1983).

In contrast, in our task we explicitly aimed to induce a state of experienced control over the outcome of the situation – by making participants believe they received online feedback on their actual task performance – and thus likely triggered participants' self-improvement motives (Jordan & Audia, 2012; Sedikides & Hepper, 2009). A situation that allows for improvement is thought to naturally trigger people's desire for self-improvement specifically in response to past failure (Taylor, Neter, & Wayment, 1995) and when upward social comparison information is provided (Sedikides & Hepper, 2009). Given the participants feel they have the necessary psychological resources (i.e. sufficient self-esteem, low self-threat induced by a novel task), they should focus on the negative feedback – here trials that, in particular, offer room for improvement – and be motivated to do better in the current performance situation (Sedikides & Hepper, 2009). The increased significance of negative feedback, that has also been shown to shape performance forecasts (Clark & Friesen, 2009; Ertac, 2011), might have led to the biased updating behavior we find in the current study. This particular study set-up aligns nicely with many real-life performance situations at school or work environments, in which negative feedback calls for direct behavioral change (Sedikides & Hepper, 2009). One needs to consider, that this interpretation is speculative in the context of our experiments and we do not know if the motivation to improve might have driven the individuals' learning behavior. While participants did believe that the relative performance feedback was related to their actual task performance, most likely believing they could change the next trial's outcome with their behavior, participants did not receive concrete feedback on their estimation accuracy and were not actually able to improve their performance. An alternative factor shaping learning biases might be the reduced relevance of the estimation task in contrast to learning about one's IQ or health risks, which might have reduced positivity biases.

An additional explanation for the negativity bias might be grounded in affective processes

associated with negative feedback. From earlier research we know that failing unexpectedly in a performance situation not only triggers the motivation to improve in the next trial, but can also elicit the experience of embarrassment (even by the mere thought or possibility of an audience witnessing one's mistakes) and might induce a fear of failing again (Leary, 1995; Müller-Pinzler et al., 2015). Such an affective connotation of negative prediction errors might thus similarly increase the subjective relevance of negative information resulting in a more negatively biased learning as it is thought to form self-efficacy beliefs (Bandura & Locke, 2003). Future studies will be needed to directly test the impact of distinctive affective states like embarrassment and motives of self-enhancement or improvement as well as task-related effects on specific biases in self-related learning.

With our second main question we assessed the impact of interindividual personality differences on self-related belief formation. Apart from a general updating bias towards negative information about the self we observed that the asymmetry in learning rates was associated with prior beliefs about the self. Individuals with more negative prior beliefs about themselves (i.e. lower self-esteem) showed more pronounced learning biases towards negative information suggesting that, besides a general valence induced bias, confirmation biases shape social learning processes. The term "confirmation bias" describes the observation that one favors information in line with one's prior views and it is argued that the confirmation bias might be especially pronounced with regards to self-concepts (Swann, 1983). Such an impact of prior beliefs is supported by previous studies showing that individuals preferably updated their beliefs in line with their prior expectations about their own task performance (Ertac, 2011). Similarly, low self-esteem has been shown to lead individuals to confirm their prior beliefs, maintaining negative performance expectations even in the context of successful performance (Blascovich & McFarlin, 1981). In this line, interindividual differences in updating behavior have been associated with distinct activation patterns on the neural systems level. For example, trait optimism has been associated with decreased activation of the right inferior prefrontal gyrus in response to negative self-related information, indicating decreased sensitivity for negative information that is incongruent with more pronounced optimistic beliefs (Sharot et al., 2011).

Apart from self-esteem, trait social anxiety is a potent modulator of biases in self-related learning. In the present studies we directly addressed the fundamental fear in social anxiety: being observed in an evaluative performance situation. In line with recent findings, individuals higher in social anxiety exhibited an increased bias towards negative information (Button et al., 2015; Koban et al., 2017). Interestingly, with regard to our third main question, this learning bias towards negative information was modulated by the social context and only present when participants were exposed to a potentially judging audience. This distinction has not been explored so far in social anxiety even though the importance of the social context has recently been pointed out for depression (Safra, Chevallier, & Palminteri, 2019). It has been suggested that diminished striatal involvement in the brain's reward system reflects the lack of a motivational preference for positive social information in social anxiety disorder (Cremers, Veer, Spinhoven, Rombouts, &

Roelofs, 2015). However, given our data we would question whether such a valence bias in social anxiety exists independently of the social context. Previous studies reported that socially anxious individuals displayed negativity biases in response to social evaluative feedback on a public speech (Koban et al., 2017) or social feedback in the form of "personal-descriptive adjectives" (Button et al., 2015), information that is likely to trigger social fear related thought and attention patterns. This is corroborated by studies suggesting that individuals suffering from social anxiety typically pay more attention to information indicating a potential threat to their social image and interpret these social cues in a negatively biased way (Alden, Taylor, Mellings, & Laposa, 2008; Amin, Foa, & Coles, 1998; Amir, Prouvost, & Kuckertz, 2012; Ashbaugh, Antony, McCabe, Schmidt, & Swinson, 2005; Heimberg et al., 2010). In a previous study, we demonstrated that socially anxious individuals paid increased attention to the audience while receiving feedback on their estimation performance. Also, pupil dwell time on the faces of the audience mediated neural activation differences in the mentalizing network (Müller-Pinzler et al., 2015). Taken together, one may assume that the informational content (i.e. social evaluative feedback) as well as context (i.e. publicity or privacy) modulate attentional and cognitive processes in social anxiety. Here, we explicitly considered the impact of content and context and designed our task to be generally unrelated to social or other specific fears (unlike other commonly employed tasks in social learning where one e.g. is being judged by others with regards to one's personality). Thus, interindividual differences in updating behavior in the present data likely reflect a more basic bias in self-related information processing compared to tasks which comprise strong priors (e.g. "I have always been awkward in social interactions."; Button et al., 2015; Koban et al., 2017). Our finding of context specific biases in self-related learning thus provides first indications that biased self-related belief updating only emerges when fear related processes are triggered, i.e. when confronting (socially anxious) individuals with potential social judgement by an evaluative audience. However, as our sample consisted of healthy individuals with non-clinical levels of social anxiety, future studies on clinical populations are needed to substantiate this clinically relevant claim.

When taking self-improvement motives into account, increased responsiveness to negative feedback related to high levels of social anxiety or low self-esteem might be a strategy to make up for perceived personal deficits. However, it has been shown that people low in self-esteem rather show a decline in task performance instead of benefitting from negative feedback (Shrauger & Rosenberg, 1970), while high self-esteem is thought to facilitate task persistence (Baumeister, Campbell, Krueger, & Vohs, 2003). Similarly, instead of improving their social performance, socially anxious individuals feel unable to make the desired positive impression, which in turn increases the experience of social anxiety (Schlenker & Leary, 1982) and the kind of avoidance behavior that contributes to the often described deficits in everyday life functioning in social anxiety disorder (Heimberg et al., 2010). Self-related confirmation biases in learning might thus, at least in some cases, confirm maladaptive and rather pessimistic views about the self – or "more realistic", as discussed in the context of depression (Korn, Sharot, Walter, Heekeren, &

Dolan, 2014; Moore & Fresco, 2012). With respect to social anxiety this might reflect a core process of clinical relevance that could explain the persistence of negative self-related beliefs in the social domain that is not necessarily grounded in actual negative feedback (Heimberg et al., 2010). In line with this, another explanation for the pronounced negativity bias in the current experiments might be that individuals have overly negative prior self-related beliefs for their estimation ability, which could lead to a confirmation bias for their own inability to solve the estimation questions. Identifying the circumstances under which a preference for negative feedback might be triggered and those that lead to self-improvement vs. consistent negative beliefs, are two interesting avenues for future research.

To summarize, our results indicate a negativity bias when forming beliefs about one's own abilities in a performance situation that is shaped by prior beliefs about the self (self-esteem) in line with a confirmation bias. With the current task we were able to form people's beliefs about their own abilities within a short period of time. Such beliefs are often considered as rather stable and form the basis for human behavior in everyday social and professional life. Being able to induce and observe self-related learning processes enables us to further disentangle the basic mechanisms underlying the persistence of (negative) self-images as well as potential behavioral consequences specifically in individuals low in self-esteem or high in social anxiety. Thus, the present findings are of high relevance for developmental, educational or clinical applications.

The LOOP task introduced here has a number of unique features that we believe are important to illuminate a wider gamut of learning situations. While past research had focused on feedback on highly valued and difficult to change aspects of the self, our task explores learning in changeable and relatively neutral domains. The online performance-feedback loop suggests that people have an opportunity to directly use feedback to improve performance and the novel and neutral content of the task reduces the impact of domain specific prior beliefs. We are curious if the observed negativity bias would hold over a variety of self-related performance tasks that suggest an immediate opportunity to improve. We believe that by challenging the generalizability of the positivity bias the current study points to the importance of situational, motivational, and interindividual factors in self-related belief formation. While overly negative distortions of self-related learning might have far-reaching consequences for decisions that are crucial for everyday life our finding might also encourage a discussion about the value of recognizing personal failures as a prerequisite for improvement. Taking into account that much of what people believe is biased or even wrong (Hilbert, 2012), such intellectual humility (Leary et al., 2017) to focus on one's shortcoming or even "stupidity" (Schwartz, 2011) have recently been coined as key components for progress in research and likely a lot of other areas of life.

2.5 Materials and Methods

Participants. The study was approved by the ethics committee of the University of Lübeck (AZ 16-315, AZ 17-220), has been conducted in compliance with the ethical guidelines of the American Psychological Association (APA), and all subjects gave written informed consent. All participants were recruited at the University Campus of Lübeck, were fluent in German, and had normal or corrected-to-normal vision. All participants received monetary compensation for their participation in the study. Across all three experiments seven subjects were excluded after participation because they did not believe the cover-story of the task. For the first experiment we initially recruited 26 participants and included 24 (12 female, aged 20-31 years; *M* = 23.75; *SD* = 3.22). For the second experiment we initially recruited 64 subjects and included 61, who were randomly assigned to either a Private or a Public social context group. The Private group consisted of 30 participants (20 female, aged 18-32 years; *M* = 22.27; *SD* = 3.01), the Public group of 31 participants (22 female, aged 19-32 years; *M* = 22.58; *SD* = 2.69). For the third experiment we initially recruited 32 participants and included 30 (24 female, aged 18-30 years; *M* = 21.70; *SD* = 3.33). For details on the sample characteristics see Supplementary Table 2.1.

General procedure. *Learning of own performance task.* The Learning of own performance (LOOP) task enables participants to incrementally learn about themselves from trial-by-trial performance feedback in a task testing their own abilities. For this purpose we adapted a cognitive estimation task that we implemented in a previous study on the induction of embarrassment (Müller-Pinzler et al., 2015). For the LOOP task all participants were invited to take part in an experiment on "cognitive estimation". Participants needed to estimate properties of different objects (e.g. the height of houses or the weight of animals). To make participants learn about their estimation ability, they received manipulated performance feedback in two distinct estimation categories. Unbeknownst to the participant, one category was arbitrarily paired with High Ability and one with Low Ability feedback (e.g. “height” of houses = High Ability and “weight” of animals = Low Ability or vice versa; estimation categories were counterbalanced between Ability conditions) independently of the actual responses given by the participants. Thus, participants could learn over the course of the experiment that they performed well in one estimation category and poorly in the other. Introducing a High and a Low Ability condition also increased the variance of positive and negative prediction errors (PEs) and allowed us to assess PE valence specific effects. Performance feedback was provided after every estimation trial during the task so that participants could use the last feedback in order to adapt their predictions of the performance feedback for the next trial of the same condition. Importantly, by implementing a continuous performance-feedback-loop participants were made to believe that they could utilize the feedback in order to improve their cognitive estimation performance, e.g. to increase their efforts following negative feedback. Fixed performance feedback sequences were presented for all participants, indicating their current estimation accuracy as percentiles compared to an alleged reference group of 350 university students who, according to the cover-story, had been

tested beforehand (e.g. "You are better than 96% of the reference participants."; see Figure 2.1A). Participants never received feedback on how close their actual performance was to the "correct" answer. Presenting estimation accuracies by means of percentiles therefore ensured that participants were more likely to believe that the feedback represented their actual performance. In the Low Ability condition, feedback was approximately normally distributed around the 35th percentile (SD ≈ 16; range 1–60%) and in the High Ability condition around the 65th percentile (SD ≈ 16; range 40–99%).
In the beginning of each trial a cue (CUE) was presented indicating the estimation category (e.g. "height", which could correspond to the High Ability condition) and participants were asked to indicate their expected performance (EXP) for this trial on the same percentile scale used for feedback. Participants were told accurate EXP ratings would be rewarded with up to 6 cents per trial, i.e. the better their EXP rating matched their actual feedback percentile the more money they would receive, to increase motivation and encourage honest response behavior. Following each EXP rating, the estimation question was presented for 10 seconds (EST). During the EST period, continuous response scales below the pictures determined a range of plausible answers for each question, and participants indicated their responses by navigating a pointer on the response scale with a computer mouse. Subsequently, feedback (FB) was presented for 5 seconds (see Figure 2.1A). All stimuli were presented using MATLAB Release 2015b (The MathWorks, Inc.) and the Psychophysics Toolbox (Brainard, 1997). The LOOP formed the main frame for all three experiments. The adaptations of the LOOP for each experiment are explained below.

Experiment 1: Agent-LOOP. For the Agent-LOOP two participants were invited at the same time. Participants were informed they would take turns with the other participant, either performing the task themselves (Self) or observing the other person performing (Other). In the beginning of each trial the CUE indicated who's turn it was (e.g. "Thomas" or "You") along with the estimation category depicted below (e.g. "height"; estimation categories were counterbalanced between Ability conditions and Agent conditions (Self vs. Other)). Depending on the corresponding condition participants then indicated their EXP rating either for their own or the other participant's performance. At the end of each trial, performance feedback was always presented to both participants. Participants thus underwent four feedback conditions with 25 trials each (Agent condition (Self vs Other) × Ability condition (High Ability vs Low Ability)). Trials of all conditions were intermixed in a fixed order with a maximum of two consecutive trials of the same condition.

Experiment 2: Audience-LOOP. In experiment 2 we implemented another version of the LOOP task to assess the impact of the presence of an audience on self-related learning in a between-subject design (Figure 2.1C). Participants were invited alone and randomly assigned to one of two experimental groups. In the Private group participants completed the estimation task as described above all on their own. In the Public group the experimenter, who represented the audience, was seated behind the participant and observed his/her performance, allegedly in order to assess additional performance

characteristics that could not be recorded by the computer. The following part of the experiment including the estimation task was executed as described above. For each of the two self-related Ability conditions (High Ability vs Low Ability) 30 trials were presented intermixed in a fixed order with a maximum of two consecutive trials of the same condition.

Experiment 3: replication and extension. In experiment 3 we used the Agent-LOOP task (experiment 1) and additionally introduced publicity in a more minimal fashion compared to the Audience-LOOP (experiment 2). To do so, instead of seating someone behind the participants, we simply manipulated the amount of information participants were able to see from each other. Thus, all participants were told that they were randomly selected for the Public group by the computer (i.e. being observed), while allegedly the other participant was in the Private condition (i.e. being the observer). Like in the Agent-LOOP in experiment 1, participants were only able to see the other participant's performance feedback, but were told that their EXP ratings were made public for the other participant. This minimal change in the paradigm was expected to make participants experience being observed by and exposed to the other's judgement, while at the same time being unable to observe and judge the other person equally. This was confirmed by our debriefing questionnaire indicating that only 9% of participants in the private Agent-LOOP were bothered by the other participant observing their performance while in the public version 38% reported the same.

Statistical analysis. *Model free behavioral analysis.* A model free analysis was performed on the participants' EXP ratings for each trial to illustrate the basic effects we see in our behavioral data. For the Agent-LOOP task (experiments 1 and 3) a repeated-measures ANOVA was calculated with the factors Trial (25 Trials) × Ability condition (High Ability vs Low Ability) × Agent condition (Self vs Other). For the Audience-LOOP (experiment 2) we calculated a repeated-measures ANOVA with the factors Trial (30 Trials) × Ability condition (High Ability vs Low Ability) and Audience (Public vs Private) as a between subject factor. Additionally, we collapsed the data of experiments 1 and 3 to replicate and extend the conclusions on the impact of the audience on learning about the self and another person. The corresponding ANOVA included Audience (Public vs Private) as an additional between-subject factor. After model fitting four subjects had to be excluded from further analyses. To keep the sample consistent across analyses, model free behavioral analyses were also conducted on the reduced sample and results remained consistent with those computed on the full sample (see 2.7 Supplementary Results).

Computational modeling of learning behavior. We modeled dynamic changes in self-related beliefs for all EXP ratings participants provided in the beginning of each trial in response to the provided performance FB using prediction error delta-rule update equations (adapted Rescorla-Wagner model; Rescorla & Wagner, 1972; see Figure 2.1B). The model space is described in the Results section and depicted in Figure 2.3. In our task, Ability condition and PE valence were correlated in the sense that the Low Ability condition contained more negative PEs and the High Ability condition more positive PEs,

assuming that participants initially expect their own performance to be around the 50th percentile. Nevertheless, if the Valence Model won it could be assumed that PE valence is the more prominent factor affecting learning rates compared to the Ability condition and vice versa.
In addition to the learning rates we either fitted parameters for the initial belief about the own and the other participant's performance, separately or combined for both ability conditions, or used the initial performance expectation ratings as fixed starting values. The models presented in the Results section included initial belief parameters for each condition separately (see 2.7 Supplementary Methods for a detailed description of the complete model space).

Model fitting. For model fitting we used the RStan package (Stan Development Team, 2016. RStan: the R interface to Stan. R package version 2.14.1.), which uses Markov chain Monte Carlo (MCMC) sampling algorithms. All of the learning models in the model space were fitted for each subject in the corresponding experimental group. Posterior parameter distributions were sampled for each subject. A total of 2400 samples were drawn after 1000 burn-in samples (overall 3400 samples; thinned with a factor of 3) in three MCMC chains. We assessed if MCMC chains converged to the target distributions by inspecting $\hat{R}$ values for all model parameters (Gelman & Rubin, 1992). Three subjects (n = 1 for each of the experiments) were excluded because at least one model parameter had $\hat{R}$ values exceeding 1.1 indicating non-convergence of the MCMC chains, which was confirmed by visual inspection. An additional subject was excluded after visual inspection due to implausible model parameters, i.e. mean learning rate of 1, which was more than 10 standard deviations above average (experiment 2). Effective sample sizes (n_{eff}) of model parameters, which are estimates of the effective number of independent draws from the posterior distribution, were typically greater than 1000 (>1300 for most parameters). Posterior distributions for all parameters for each of the subjects were summarized by their mean as the central tendency resulting in a single parameter value per subject that we used in order to calculate group statistics. Using the median lead to similar conclusions.

Bayesian model selection and family inference. In order to select the model that most likely guided the participants' updating behavior, as a first step, we estimated pointwise out-of-sample prediction accuracy for all fitted models separately for each participant by approximating leave-one-out cross-validation (LOO; i.e. corresponding to leave-one-trail-out per subject) as recommended for assessing model fit without introducing penalties for model complexity (Acerbi, Dokka, Angelaki, & Ma, 2018; Gelman, Hwang, & Vehtari, 2014). To do so we applied Pareto-smoothed importance sampling (PSIS) using the log-likelihood calculated from the posterior simulations of the parameter values as implemented by Vehtari et al. (2017). Sum PSIS-LOO scores for each model as well as information about $\hat{k}$ values – the estimated shape parameters of the generalized Pareto distribution – indicating the reliability of the PSIS-LOO estimate are depicted in Table 2.1.

As summarized in Table 2.1 very few trials resulted in insufficient parameter values for $\hat{k}$ and thus potentially unreliable PSIS-LOO scores (on average 0.17 trials per subject with $\hat{k} > 0.7$, Vehtari et al., 2017). Visual inspection of the corresponding subjects suggested that in some cases subjects had provided EXP ratings far away from the current average, PSIS-LOO scores for the corresponding trials were, however, mostly within the range of the other trials. In order to make sure that these trials would not bias the model selection processes, we excluded the PSIS-LOO scores for these trials and repeated the model selection procedure replicating our model selection results. Bayesian model selection (BMS) on PSIS-LOO scores was performed on the group level accounting for group heterogeneity in the model that best describes learning behavior (Rigoux, Stephan, Friston, & Daunizeau, 2014). This procedure provides the protected exceedance probability for each model (*pxp*), indicating how likely a given model has a higher probability explaining the data than all other models in the comparison set, as well as the Bayesian omnibus risk (*BOR*), the posterior probability that model frequencies for all models are all equal to each other (Rigoux et al., 2014). We also provide difference scores of PSIS-LOO in contrast to the model that won the BMS that can be interpreted as a simple 'fixed-effect' model comparison (see Table 2.1 and Supplementary Tables 2.2 - 2.3; Acerbi et al., 2018; Vehtari et al., 2017). Mostly, model comparisons according to PSIS-LOO difference scores were qualitatively comparable to the BMS analyses for our data.

Posterior predictive checks and statistical analyses of learning parameters. First, posterior predictive checks were conducted by quantifying if the predicted data could capture the variance in EXP ratings for each subject within each of the experimental conditions using Regression analyses. Additionally, we repeated the model free analysis we had done on the behavioral data with the data predicted by the winning model to assess if the winning model captured the core effects in the behavioral data. Additionally, correlations between the parameters within the winning model were assessed (see 2.7 Supplementary Results and Supplementary Tables 2.4 - 2.6).

Model parameters, i.e. learning rates, of the winning models for all experiments were analyzed on the group level using IBM SPSS Statistics for Windows, Version 22.0 (IBM Corp., 2013, Armonk, NY). For the Agent-LOOP in experiment 1 a repeated-measures ANOVA was calculated on the learning rates with the factor Agent (Self [$\alpha_{PE+(S)}$, $\alpha_{PE-(S)}$] vs Other [$\alpha_{PE+(S)}$, $\alpha_{PE-(S)}$]) and factor PE Valence (PE+ [$\alpha_{PE+(S)}$, $\alpha_{PE+(O)}$] vs PE− [$\alpha_{PE-(S)}$, $\alpha_{PE-(O)}$]) testing if negative feedback gains a specific weight when learning about the self vs the other.

In the Audience-LOOP we assessed the impact of the social context, i.e. the presence of an evaluative audience, on self-related belief updating and its interaction with social anxiety. Here, an ANOVA was implemented with PE Valence (PE+ [α_{PE+}] vs PE- [α_{PE-}]) as a within-subject factor and Audience (Public vs Private) as a between subject factor.

For experiment 3, as for the Agent-LOOP in experiment 1, a repeated-measures ANOVA was calculated on the learning rates with the factors Agent and PE Valence. Additionally,

we collapsed the learning rates of experiments 1 and 3 to directly test the impact of the audience on learning about the self and another person. We thus implemented another ANOVA including Audience (Public vs Private) as an additional between-subject factor.
To investigate the associations of learning biases with the subjective prior sense of self-esteem, i.e. SDQ-III scores (available for experiment 1 and 3), as well as social anxiety, i.e. SIAS scores, we calculated a normalized learning rate Valence Bias Score for self-related learning (Valence Bias Score = $(\alpha_{PE+(S)} - \alpha_{PE-(S)})/(\alpha_{PE+(S)} + \alpha_{PE-(S)})$) and similarly for other-related learning (Niv et al., 2012; Palminteri et al., 2017). Pearson correlations were calculated between Valence Bias Score and personality traits. Context specific effects of social interaction anxiety on self- vs other-related updating behavior were assessed by contrasting correlations between the public and private groups. For all three experiments we additionally tested if the Valence Bias Score was suitable to capture interindividual differences in how subjects changed their beliefs about themselves over time by calculating partial correlations between Valence Bias Scores and the average of the last two EXP ratings for both ability conditions controlling for the average of the first two EXP ratings for both ability conditions.
Finally, cumulative Bayesian analyses (using JASP Version 0.9, ASP Team, 2018) were implemented collapsing the data for all experiments in order to assess the overall evidence for self-related learning biases as well as associations of such learning biases with social anxiety and self-esteem. Here, we first assessed the impact of the audience on each of the effects. A Bayesian ANOVA with the factors PE Valence and Audience was thus calculated on the learning rates and Bayesian linear regressions of personality traits and Valence Bias Scores were calculated including the factor Audience and the interaction of Audience and personality traits. Effect sizes for self-related learning biases were then calculated using Bayesian t-tests on the learning rates and Bayesian correlations were calculated to assess the associations of learning biases with personality traits. In case there was evidence for an audience effect, effect sizes were calculated separately for the Public and the Private group.

2.6 References

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. In S. J. Gershman (Ed.), *PLoS Computational Biology* (Vol. 14). https://doi.org/10.1371/journal.pcbi.1006110

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. In S. J. Gershman (Ed.), *PLoS Computational Biology* (Vol. 14). https://doi.org/10.1371/journal.pcbi.1006110

Alden, L. E., Taylor, C. T., Mellings, T. M. J. B., & Laposa, J. M. (2008). Social anxiety and the interpretation of positive social events. *Journal of Anxiety Disorders*, *22*(4), 577–590. https://doi.org/10.1016/j.janxdis.2007.05.007

Amin, N., Foa, E. B., & Coles, M. E. (1998). Negative interpretation bias in social phobia. *Behaviour Research and Therapy*, *36*(10), 945–957. https://doi.org/10.1016/S0005-7967(98)00060-6

Amir, N., Prouvost, C., & Kuckertz, J. M. (2012). Lack of a Benign Interpretation Bias in Social Anxiety Disorder. *Cognitive Behaviour Therapy*, *41*(2), 119–129. https://doi.org/10.1080/16506073.2012.662655

Ashbaugh, A. R., Antony, M. M., McCabe, R. E., Schmidt, L. A., & Swinson, R. P. (2005). Self-evaluative biases in social anxiety. *Cognitive Therapy and Research*, *29*(4), 387–398. https://doi.org/10.1007/s10608-005-2413-9

Bandura, A. (2001). Social Cognitive Theory: An Agentic Perspective. *Annual Review of Psychology*, *52*(1), 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-Efficacy Beliefs as Shapers of Children's Aspirations and Career Trajectories. *Child Development*, *72*(1), 187–206. https://doi.org/10.1111/1467-8624.00273

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, *88*(1), 87–99. https://doi.org/10.1037/0021-9010.88.1.87

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does High Self- Esteem Cause Better Performance, Interpersonal Success, Happiness or Healtier Lifestyles? *Psychological Science in the Public Interest*, *4*(1), 1–44.

Bem, D. J. (1965). An experimental analysis of self-persuasion. *Journal of Experimental Social Psychology*, *1*(3), 199–218. https://doi.org/10.1016/0022-1031(65)90026-0

Blascovich, J., & McFarlin, D. B. (1981). Effects of self-esteem and performance feedback on future affective preferences and cognitive expectations. *Journal of Personality and Social Psychology*, *40*(3), 521–531. https://doi.org/10.1037/0022-3514.40.3.521

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Brown, J. D. (1986). Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments. *Social Cognition*, *4*(4), 353–376. https://doi.org/10.1521/soco.1986.4.4.353

Brown, J. D. (2012). Understanding the Better Than Average Effect. *Personality and Social Psychology Bulletin*, *38*(2), 209–219. https://doi.org/10.1177/0146167211432763

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*(4), 676–713. https://doi.org/10.1037/0033-295X.106.4.676

Button, K. S., Kounali, D., Stapinski, L., Rapee, R. M., Lewis, G., & Munafò, M. R. (2015). Fear of negative evaluation biases social evaluation inference: Evidence from a probabilistic learning task. *PLoS ONE*, *10*(4), 1–15. https://doi.org/10.1371/journal.pone.0119456

Clark, J., & Friesen, L. (2009). Overconfidence in forecasting of own performance: an experimental study. *Economic Journal*, *119*(2004), 229–251. https://doi.org/10.1111/j.1468-0297.2008.02211.x

Cremers, H. R., Veer, I. M., Spinhoven, P., Rombouts, S. A. R. B., & Roelofs, K. (2015). Neural sensitivity to social reward and punishment anticipation in social anxiety disorder. *Frontiers in Behavioral Neuroscience*, *8*(January), 1–9. https://doi.org/10.3389/fnbeh.2014.00439

Eccles, J. S. (1989). Bringing Young Women to Math and Science. In *Gender and Thought: Psychological Perspectives* (pp. 36–58). https://doi.org/10.1007/978-1-4612-3588-0_3

Eil, D., & Rao, J. M. (2010). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior and Organization*, *80*(3), 532–545. https://doi.org/10.1016/j.jebo.2011.05.012

Garner, M., Mogg, K., & Bradley, B. P. (2006). Fear-relevant selective associations and social anxiety: Absence of a positive bias. *Behaviour Research and Therapy*, *44*(2), 201–217. https://doi.org/10.1016/j.brat.2004.12.007

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Goldin, P. R., Manber-Ball, T., Werner, K., Heimberg, R., & Gross, J. J. (2009). Neural Mechanisms of Cognitive Reappraisal of Negative Self-Beliefs in Social Anxiety Disorder. *Biological Psychiatry*, *66*(12), 1091–1099. https://doi.org/10.1016/j.biopsych.2009.07.014

Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2010). A cognitive- behavioral model of social anxiety disorder: Update and extension. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social anxiety: Clinical, developmental, and social perspectives* (pp. 395–422). New York: NY: Elsevier.

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211–237. https://doi.org/10.1037/a0025940

Hirsch, C. R., & Mathews, A. (2000). Impaired positive inferential bias in social phobia. *Journal of Abnormal Psychology*, *109*(4), 705–712. https://doi.org/10.1037/0021-843X.109.4.705

Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). Oxford.

Jordan, A. H., & Audia, P. G. (2012). Self-Enhancement and Learning from Performance Feedback. *Academy of Management Review*, *37*(2), 211–231. https://doi.org/10.5465/amr.2010.0108

Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037//0033-2909.119.2.254

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., … Arch, J. J. (2017). Social Anxiety is Characterized by Biased Learning About Performance and the Self. *Emotion*. https://doi.org/10.1037/emo0000296

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Korn, Christoph W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592. https://doi.org/10.1017/S0033291713001074

Krueger, N., & Dickson, P. R. (1994). How Believing in Ourselves Increases Risk Taking: Perceived Self-Efficacy and Opportunity Recognition. *Decision Sciences*, *25*(3), 385–400. https://doi.org/10.1111/j.1540-5915.1994.tb00810.x

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162. https://doi.org/10.1016/j.neuroimage.2016.02.014

Leary, M. R. (1995). *Self-presentation: Impression management and interpersonal behavior*. Madison, WI: Brown & Benchmark Publishers.

Leary, M. R. (2007). Motivational and Emotional Aspects of the Self. *Annual Review of Psychology*, *58*(1), 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Leary, M. R., & Atherton, S. C. (1986). Self-Efficacy, Social Anxiety, and Inhibition in Interpersonal Encounters. *Journal of Social and Clinical Psychology*, *4*(3), 256–267. https://doi.org/10.1521/jscp.1986.4.3.256

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., … Hoyle, R. H. (2017). Cognitive and Interpersonal Features of Intellectual Humility. *Personality and Social Psychology Bulletin*, *43*(6), 793–813. https://doi.org/10.1177/0146167217697695

Leary, M. R., & Kowalski, R. M. (1995). *Social Anxiety*. New York: The Guilford Press.

Loewenstein, G. (2006). The Pleasures and Pains of Information. *Science*, *312*(5774), 704–706. https://doi.org/10.1126/science.1128388

Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology: General*, *105*(1), 3–46. https://doi.org/10.1037/0096-3445.105.1.3

Markus, H. R., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology*, *38*(1), 299–337. https://doi.org/10.1146/annurev.psych.38.1.299

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The Construct Validity of Multidimensional Self-Concept Ratings by Late Adolescents. *Journal of Educational Measurement*, *21*(2), 153–174.

Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy*, *36*(4), 455–470.

Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. (2013). Managing Self-Confidence: Theory and Experimental Evidence. *Ssrn*. https://doi.org/10.2139/ssrn.2285056

Moore, M. T., & Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, *32*(6), 496–509. https://doi.org/10.1016/J.CPR.2012.05.004

Morrison, A. S., & Heimberg, R. G. (2013). Social Anxiety and Social Anxiety Disorder. *Ssrn*, *9*, 1029–1036. https://doi.org/10.1146/annurev-clinpsy-050212-185631

Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., … Krach, S. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, *119*(0), 252–261. https://doi.org/10.1016/j.neuroimage.2015.06.036

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/jneurosci.5498-10.2012

Nolen-Hoeksema, S., Girgus, J. S., & Seligman, M. E. (1986). Learned helplessness in children: A longitudinal study of depression, achievement, and explanatory style. *Journal of Personality and Social Psychology*, *51*(2), 435–442. https://doi.org/10.1037/0022-3514.51.2.435

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement. In A. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Safra, L., Chevallier, C., & Palminteri, S. (2019). Depressive symptoms are associated with blunted reward learning in social contexts. *PLOS Computational Biology*, *15*(7), e1007224. https://doi.org/10.1371/journal.pcbi.1007224

Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: A conceptualization model. *Psychological Bulletin*, *92*(3), 641–669. https://doi.org/10.1037/0033-2909.92.3.641

Schwartz, M. A. (2011). the Importance of Stupidity in Scientific Research. *Seismological Research Letters*, *82*(1), 3–4. https://doi.org/10.1785/gssrl.82.1.3

Sedikides, C., & Gregg, A. P. (2008). *Self-Enhancement Food for Thought*. *3*(2), 102–116.

Sedikides, C., & Hepper, E. G. D. (2009). Self-Improvement. *Social and Personality Psychology Compass*, *3*, 899–917. https://doi.org/10.1111/j.1751-9004.2009.00231.x

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/j.tics.2015.11.002

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. https://doi.org/10.1038/nn.2949

Shrauger, J. S., & Rosenberg, S. E. (1970). Self-esteem and the effects of success and failure feedback on performance. *Journal of Personality*, *38*(3), 404–417. https://doi.org/10.1111/j.1467-6494.1970.tb00018.x

Steinmetz, J., Xu, Q., Fishbach, A., Zhang, Y., Xu, Q., & Steinmetz, J. (2016). Being Observed Magnifies Action Janina. *Journal of Personality and Social Psychology*, *111*(6), 852–865. https://doi.org/10.1017/CBO9781107415324.004

Swann, W. B. (1983). Self-verification: Bringing social reality into harmony with the self. In J. Suls & A. G. Greenwald (Eds.), *Social psychological perspectives on the self* (Vol. 2, pp. 33–66). https://doi.org/10.1126/science.218.4574.782

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210.

Taylor, S. E., Neter, E., & Wayment, H. A. (1995). Self-Evaluation Processes. *Personality and Social Psychology Bulletin*, *21*(12), 1278–1287. https://doi.org/10.1177/01461672952112005

Triplett, N. (1898). The Dynamogenic Factors in Pacemaking and Competition. *The American Journal of Psychology*, *9*(4), 507–533.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vroling, M. S., & De Jong, P. J. (2009). Deductive reasoning and social anxiety: Evidence for a fear-confirming belief bias. *Cognitive Therapy and Research*, *33*(6), 633–644. https://doi.org/10.1007/s10608-008-9220-z

Zajonc, R. B., & Sales, S. M. (1966). Social facilitation of dominant and subordinate responses. *Journal of Experimental Social Psychology*, *2*(2), 160–168. https://doi.org/10.1016/0022-1031(66)90077-1

Zimmerman, B. J. (1990). Self-Regulated Learning and Academic Achievement: An Overview. *Educational Psychologist*, *25*(1), 3–17. https://doi.org/10.1207/s15326985ep2501_2

2.7 Supplementary Information

Supplementary Methods

Further information on the experimental procedure.

For the Agent-LOOP (experiments 1 and 3) both participants arrived at the same time, were led to the same room and instructed about the task by the experimenter. After signing the informed consent forms, they went to separate rooms to fill out demographic and self-esteem questionnaires and to practice the estimation task. Each room was equipped with a desktop computer and participants were told that the two computers were connected. While the other participant completed the estimation question, participants could see the estimation question, but not the answer given by the other participant. For the private version of the Agent-LOOP the EXP rating was private as well and could not be seen by the other participant.

For all three experiments participants were asked to fill in personality questionnaires and a post-experimental questionnaire. Before leaving, all participants received compensation for their participation, including an additional 6 cents per trial promised for "accurate" EXP ratings, and were debriefed about the study. The total duration of the experiments, including post-experiment questionnaires, was 1.5 to 2 hours.

The sample recruited for experiment 2 was part of another project assessing the impact of stress on self-related learning. All participants therefore completed a simple reading task prior to the estimation task, which served as a control condition for a stress task. This took approximately 45 minutes and during the whole task period cortisol samples were collected.

Detailed information on the model space.

To see whether a learning model can capture the participants' behavior and allows us to summarize the data using principled parameters such as learning rates, we performed a model comparison (see Figure 2.3). Our model space contained three main models varying with regards to their assumptions about biased updating behavior when learning about the self (see Figure 2.3). The simplest learning model used one single learning rate for the whole behavioral time course for each participant, thus not assuming any learning biases [$EXP_{t+1} = EXP_t + \alpha_{Uni}\ PE_t$, while $PE_t = FB_t - EXP_t$; Unity Model]. The second model, the Ability Model, contained a separate learning rate for each of the ability conditions, assuming that participants would show different updating behavior in the High Ability condition (α_{HA}) vs Low Ability condition (α_{LA}). The third model, the Valence Model, included separate learning rates for positive PEs (α_{PE+}) vs negative PEs (α_{PE-}) across both ability conditions, thus suggesting that the valence (positive vs negative) of the PE biases self-related learning rather than the ability condition itself. In the Agent-LOOP task (experiments 1 and 3) the distinction between learning about oneself vs another person was introduced as a second factor in the model space resulting in three additional models. Model 4 corresponded to the Unity Model with separate learning rates for the self ($\alpha_{Uni(S)}$) and the other person ($\alpha_{Uni(O)}$). Model 5 was the extension of the Ability Model distinguishing between learning about the self ($\alpha_{HA(S)}$, $\alpha_{LA(S)}$) and the other person ($\alpha_{HA(O)}$, $\alpha_{LA(O)}$), resulting

in four different learning rates. Model 6 extended the Valence Model by separate learning rates for oneself ($\alpha_{PE+(S)}$, $\alpha_{PE-(S)}$) and the other person ($\alpha_{PE+(O)}$, $\alpha_{PE-(O)}$).

All models in the complete model space not only differed with regard to the learning rates as described in the main manuscript but also with regard to the initial belief about the own and the other participant's performance. In addition to the learning rates for all 3/ 6 learning models described in the main manuscript we either fitted parameters for the initial belief about the own and the other participant's performance, separately (Models 1-3 for the Audience-LOOP and Models 1-6 for the Agent-LOOP) or combined for both ability conditions (Models 5-7 for the Audience-LOOP and Models 8-13 for the Agent-LOOP), or used the initial performance expectation ratings as fixed starting values (Models 8-10 for the Audience-LOOP and Models 14-19 for the Agent-LOOP), resulting in 18 learning models for experiments 1 and 3, and 9 learning models for experiment 2.
To test if the participants' EXP ratings could be better explained in terms of prediction error learning as compared to stable assumptions in each ability condition, we included a simple Mean Model with a mean value for each task condition. The Mean Models were numbered Model 4 for the Audience-LOOP und Model 7 for the Agent-LOOP to keep the numbering in the main manuscript consitent. For the Audience-LOOP one mean value for the High Ability condition was estimated and one for the Low Ability condition. For the Agent-LOOP four mean values were estimated for the Agent (Self vs Other) x Ability condition (High vs Low). PSIS-LOO scores for all models are reported in Supplementary Tables 2.2 and 2.3.

Supplementary Results

Additional model free behavioral analyses.
The combined analysis of the public and private Agent-LOOP (experiment 1 and 3) confirmed the results of the Audience-LOOP by showing that Audience did not have any significant effects also with regards to the additional Agent condition (main effect of Audience: $F_{(1,50)} = 0.61$, $p = .439$; Audience x Ability condition: $F_{(1,50)} = 0.53$, $p = .472$; Audience x Agent condition: $F_{(1,50)} = 0.31$, $p = .581$; Audience x Ability condition x Agent condition: $F_{(1,50)} = 0.61$, $p = .440$; Audience x Ability condition x Trial: $F_{(24,1200)} = 0.47$, $p = .987$; Audience x Agent condition x Trial: $F_{(24,1200)} = 0.76$, $p = .794$; Audience x Ability condition x Agent condition x Trial: $F_{(24,1200)} = 0.65$, $p = .903$). The remaining effects stayed consistent with the separate analyses of experiment 1 (main effect Ability condition: $F_{(1,50)} = 386.16$, $p < .001$; Trial x Ability condition: $F_{(24,1200)} = 66.55$, $p < .001$; main effect of Agent condition: $F_{(1,50)} = 32.09$, $p < .001$; Agent condition x Ability condition: $F_{(1,50)} = 6.97$, $p = .011$; Trial x Agent condition x Ability condition interaction: $F_{(24,1200)} = 1.43$, $p = .083$).

Model free behavioral analyses on the extended sample.
We replicated the model free behavioral analysis on the EXP ratings including the four participants that had been excluded after model fitting and the results matched the results of the smaller sample. For the Agent-LOOP in experiment 1 the Trial x Ability condition x Agent condition ANOVA again revealed a significant main effect of Ability condition ($F_{(1,23)} = 225.68$, $p < .001$) and Agent condition ($F_{(1,23)} = 17.60$, $p < .001$) and interactions of Trial

x Ability condition ($F_{(24,552)}$ = 31.12, p < .001) and Agent condition x Ability condition ($F_{(1,23)}$ = 5.42, p = .029). For the Audience-LOOP (experiment 2) we also found a significant main effect of Ability condition ($F_{(1,59)}$ = 262.9, p < .001) and interaction of Trial x Ability condition ($F_{(29,1711)}$ = 42.66, p < .001), while there was no significant impact of the Audience on EXP ratings (main effect of Audience: $F_{(1,59)}$ = 0.20, p = .655; Audience x Ability condition: $F_{(1,59)}$ = 0.32, p = .571; Audience x Ability condition x Trial: $F_{(29,1711)}$ = 1.08, p = .357). For the public version of the Agent-LOOP in experiment 3 the results including all participants also stayed the same (main effect of Ability condition: $F_{(1,29)}$ = 194.68, p < .001; Agent condition: $F_{(1,29)}$ = 20.24, p < .001; interaction of Trial x Ability condition: $F_{(24,672)}$ = 39.92, p < .001; no significant interaction of Agent condition x Ability condition: $F_{(1,29)}$ = 2.19, p = .149). The results of the combined analysis of the public and private Agent-LOOP (experiment 1 and 3) stayed the same as well (main effect Ability condition: $F_{(1,52)}$ = 406.92, p < .001; Trial x Ability condition: $F_{(24,1248)}$ = 69.02, p < .001; main effect of Agent condition: $F_{(1,52)}$ = 35.52, p < .001; Agent condition x Ability condition: $F_{(1,52)}$ = 7.72, p = .008; Trial x Agent condition x Ability condition interaction: $F_{(24,1248)}$ = 1.34, p = .126; Audience related effects: main effect of Audience: $F_{(1,52)}$ = 0.61, p = .439; Audience x Ability condition: $F_{(1,52)}$ = 0.27, p = .607; Audience x Agent condition: $F_{(1,52)}$ = 0.49, p = .488; Audience x Ability condition x Agent condition: $F_{(1,52)}$ = 0.82, p = .370; Audience x Ability condition x Trial: $F_{(24,1248)}$ = 0.52, p = .975; Audience x Agent condition x Trial: $F_{(24,1248)}$ = 0.72, p = .836; Audience x Ability condition x Agent condition x Trial: $F_{(24,1248)}$ = 0.73, p = .825).

Posterior predictive checks: Behavioral analyses on the predicted data.

To assess whether our winning model captured the core effects in our model free analysis, we let the parametrized winning model predict the time course of EXP for each participant, and compared these model predictions against the actual data (see Figure 2.2). Figure 2.2 visually confirms the ability of the model to capture the observed data despite its small number of parameters. We repeated the behavioral analyses we had done on the actual behavioral data on the predicted data. For the Agent-LOOP in experiment 1 the Trial x Ability condition x Agent condition ANOVA on the predicted data revealed a significant main effect of Ability condition ($F_{(1,22)}$ = 273.80, p < .001) and an interaction of Trial x Ability condition ($F_{(24,528)}$ = 199.29, p < .001). Again, there was no significant interaction of Trials x Agent condition x Ability condition ($F_{(24,528)}$ = 0.72, p = .838) but the main effect of Agent condition ($F_{(1,22)}$ = 17.09, p < .001) and the Agent condition x Ability condition interaction ($F_{(1,22)}$ = 5.40, p = .030) confirmed the negative bias for self- vs other-related evaluations we found in the behavioral data. For the Audience-LOOP (experiment 2) we also found a significant main effect of Ability condition ($F_{(1,57)}$ = 305.45, p < .001) and interaction of Trial x Ability condition ($F_{(29,1653)}$ = 210.43, p < .001) indicating that the participants' belief updating was reflected in the predicted data, while there was no significant impact of the Audience on EXP ratings (main effect of Audience: $F_{(1,57)}$ = 0.07, p = .789; Audience x Ability condition: $F_{(1,57)}$ = 0.62, p = .436; Audience x Ability condition x Trial: $F_{(29,1653)}$ = 1.23, p = .187). For the public version of the Agent-LOOP in experiment 3 the results of the predicted data were also comparable to the actual behavioral data (main effect of Ability condition: $F_{(1,28)}$ = 222.63, p < .001; interaction of Trial x Ability condition: $F_{(24,672)}$ = 227.63,

p < .001; main effect of Agent condition: $F_{(1,28)}$ = 20.02, p < .001; no significant interaction of Agent condition x Ability condition: $F_{(1,28)}$ = 1.19, p = .285; and no Trial x Agent condition x Ability condition: $F_{(24,528)}$ = 0.06, p>.999). Repeating the behavioral analysis we had done on the model free data onto the predictions thus confirmed that it recapitulates the tendency towards more negative performance expectations for the other that was core to our data.

Additional results on learning rates and parameter correlations.

Across all experiments learning rates were significantly greater than zero (all *ps* < .001) indicating that, as intended, participants updated their self-related expectations according to the provided feedback (see also Figure 2.4). This is supported by model comparisons favoring the learning models over the Mean Model that does not assume prediction error learning during the experiments (see Supplementary Tables 2.2 and 2.3).

We calculated Pearson correlations between the parameters within the winning model and for each of the experiments and found no significant correlations between learning rates and starting values (p>.05, with p-values corrected for multiple comparisons: p = .05/28 = .002 (Agent-LOOP; experiment 1/ 3) and p = .05/6 = .008 (Audience-LOOP). This suggests that learning rates were neither strongly biased nor restricted by participants' estimated EXP starting level and starting levels alone are unlikely to explain interindividual differences in learning behavior. Learning rates were positively correlated within the winning model, specifically learning rates within the Self and the Other condition (see Supplementary Tables 2.4-2.6), which makes it unlikely that differences in learning rates between positive and negative PEs would be induced by anti-correlations induced by the model fitting procedure.

Supplementary Tables

Supplementary Table 2.1. Sample characteristics

	Experiment 1		Experiment 2				Experiment 3	
			Private		Public			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	23.75	3.22	22.24	3.05	22.58	2.69	21.70	3.33
Self-esteem	6.09	0.89					6.01	0.89
SIAS	2.04	0.47	1.97	0.55	2.01	0.49	2.01	0.52

Note. Sample characteristics for the three experiments. SD = standard deviation; SIAS = averaged score on the Social Interaction Anxiety Scale; Experiment 1: N=24; Experiment 2: N(Private)=30; N(Public)=31; Experiment 3: N=30.

Supplementary Table 2.2. Model comparisons for the Agent-LOOP task

Model	LOO	LOO-SE	LOO-Diff (SE-Diff)	% of $\hat{k} > 0.7$	No. Est. Parameters
Learning Models					
Estimated IV for Self vs Other x High vs Low Ability					
Self = Other					
Unity Model (M1)	-2380.1	247.8	135.4 (63.7)	0.1	5
Ability Model (M2)	-2336.5	261.5	91.7 (42.4)	0.3	6
Valence Model (M3)	-2320.5	259.0	75.7 (49.4)	0.2	6
Self ≠ Other					
Unity Model (M4)	-2376.2	254.8	131.5 (54.6)	0.4	6
Ability Model (M5)	-2330.7	263.3	85.9 (42.8)	1.2	8
Valence Model (M6)	-2244.8	283.5	-	0.3	8
Estimated IV for Self vs Other across ability conditions					
Self = Other					
Unity Model (M8)	-2516.4	237.8	271.6 (88.5)	0.0	3
Ability Model (M9)	-2409.8	241.9	165.0 (69.6)	0.1	4
Valence Model (M10)	-2428.1	244.6	183.3 (73.0)	0.0	4
Self ≠ Other					
Unity Model (M11)	-2440.9	237.0	196.1 (87.5)	0.1	4
Ability Model (M12)	-2363.5	250.3	118.8 (63.3)	0.8	6
Valence Model (M13)	-2232.7	261.7	-12.1 (32.2)	0.1	6
Fixed IV					
Self = Other					
Unity Model (M14)	-2943.3	244.4	698.5 (126.4)	0.0	1
Ability Model (M15)	-2763.9	240.5	519.1 (106.9)	0.1	2
Valence Model (M16)	-2765.2	247.9	520.5 (101.5)	0.0	2
Self ≠ Other					
Unity Model (M17)	-2840.0	249.4	595.3 (109.8)	0.1	2
Ability Model (M18)	-2566.4	251.5	321.7 (78.3)	0.6	4
Valence Model (M19)	-2453.8	270.3	209.0 (51.7)	0.1	4
No Learning					
Mean Model (M7)	-2953.6	190.3	708.9 (123.3)	0.0	4

Note. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (Valence Model) and standard errors of differences; percentage of $\hat{k}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al. 2017); No. Est. Parameters = number of estimated parameters in the model. IV = initial parameter values for the performance expectations.

Supplementary Table 2.3. Model comparisons for the Audience-LOOP task

Model	PSIS-LOO	LOO-SE	LOO-Diff (SE-Diff)	% of $\hat{k}$>0.7	No. Est. Parameters
Learning Models					
Estimated IV for High vs Low Ability					
Unity Model (M1)	-708.2	145.1	213.1 (35.8)	0.1	3
Ability Model (M2)	-570.2	150.0	75.0 (26.8)	0.3	4
Valence Model (M3)	-495.2	150.9	-	0.1	4
Estimated IV across ability conditions					
Unity Model (M5)	-943.0	134.6	447.9 (79.9)	0.0	2
Ability Model (M6)	-733.6	141.4	238.5 (57.7)	0.2	3
Valence Model (M7)	-623.6	142.9	128.5 (61.2)	0.1	3
Fixed IV					
Unity Model (M8)	-1387.9	177.5	892.8 (149.6)	0.0	1
Ability Model (M9)	-1177.8	206.4	682.7 (170.9)	0.7	2
Valence Model (M10)	-975.8	163.8	480.6 (107.4)	0.0	2
No Learning					
Mean Model (M4)	-1189.5	124.9	694.4 (61.3)	0.0	2

Note. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (Valence Model) and standard errors of differences; percentage of $\hat{k}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al. 2017); No. Est. Parameters = number of estimated parameters in the model. IV = initial parameter values for the performance expectations.

Supplementary Table 2.4. Parameter Correlations for Experiment 1

Model Parameter	IV_{S+}	IV_{S-}	IV_{O+}	IV_{O-}	$\alpha_{PE+(S)}$	$\alpha_{PE-(S)}$	$\alpha_{PE+(O)}$	$\alpha_{PE-(O)}$
IV_{S+}			.31	.12	.26	.06	.12	.08
IV_{S-}	.40		.11	.31	.20	.20	.15	.17
IV_{O+}	.31	.11		-.02	.11	.06	.16	.33
IV_{O-}	.12	.31	-.02		.00	-.02	-.15	-.11
$\alpha_{PE+(S)}$	.26	.20	.11	.00		.67*	.71*	.74*
$\alpha_{PE-(S)}$	.06	.20	.06	-.02	.67*		.62*	.82*
$\alpha_{PE+(O)}$	.12	.15	.16	-.15	.71*	.62*		.84*
$\alpha_{PE-(O)}$	.08	.17	.33	-.11	.74*	.82*	.84*	

Supplementary Table 2.5. Parameter Correlations for Experiment 2

		Model Parameter								
		Private					Public			
		IV_+	IV_-	$\alpha_{PE+(S)}$	$\alpha_{PE-(S)}$		IV_+	IV_-	$\alpha_{PE+(S)}$	$\alpha_{PE-(S)}$
Model Parameter	IV_+		.64*	.35	.14	IV_+		.30	.02	-.14
	IV_-	.64*		.30	.21	IV_-	.30		.07	-.05
	α_{PE+}	.35	.30		.66*	α_{PE+}	.02	.07		.72*
	α_{PE-}	.14	.21	.66*		α_{PE-}	-.14	-.05	.72*	

Supplementary Table 2.6. Parameter Correlations for Experiment 3

		Model Parameter							
		IV_{S+}	IV_{S-}	IV_{O+}	IV_{O-}	$\alpha_{PE+(S)}$	$\alpha_{PE-(S)}$	$\alpha_{PE+(O)}$	$\alpha_{PE-(O)}$
Model Parameter	IV_{S+}		.04	.21	-.15	-.27	-.36	.14	-.02
	IV_{S-}	.04		-.37	-.15	.00	-.21	-.21	-.21
	IV_{O+}	.21	-.37		-.05	-.13	.19	-.16	-.19
	IV_{O-}	-.15	-.15	-.05		.03	-.07	-.07	.16
	$\alpha_{PE+(S)}$	-.27	.00	-.13	.03		.65*	.49	.46
	$\alpha_{PE-(S)}$	-.36	-.21	.19	-.07	.65*		.30	.31
	$\alpha_{PE+(O)}$	.14	-.21	-.16	-.07	.49	.30		.84*
	$\alpha_{PE-(O)}$	-.02	-.21	-.19	.16	.46	.31	.84*	

Note. Parameter correlations for the three experiments. The correlation tables include the initial parameter values for the performance expectations (IV) in the high ability condition (IV_+) and for the low ability condition (IV_-) and learning rates (α) for positive prediction errors (α_{PE+}) and for negative prediction errors (α_{PE-}). For experiments 1 and 3, as depicted in Table S4 and S6, parameters are separated for the Self condition (IV_{S+}, IV_{S-}, $\alpha_{PE+(S)}$, $\alpha_{PE-(S)}$) and the Other condition (IV_{O+}, IV_{O-}, $\alpha_{PE+(O)}$, $\alpha_{PE-(O)}$). * indicates correlations with p-values < .05, corrected for multiple comparison as described above.

3 Study 2

Neurocomputational mechanisms of affected beliefs[2]

3.1 Abstract

The feedback people receive on their behavior shapes the process of belief formation and self-efficacy in mastering a particular task. However, the neural and computational mechanisms of how the subjective value of self-efficacy beliefs, and the corresponding affect, influence the learning process remain unclear. We investigated these mechanisms during self-efficacy belief formation using fMRI, pupillometry, and computational modeling, and by analyzing individual differences in affective experience. Biases in the formation of self-efficacy beliefs were associated with affect, pupil dilation, and neural activity within the anterior insula, amygdala, ventral tegmental area/ substantia nigra, and mPFC. Specifically, neural and pupil responses mapped the valence of the prediction errors in correspondence with individuals' experienced affective states and learning biases during self-efficacy belief formation. Together with the functional connectivity dynamics of the anterior insula within this network, our results provide evidence for neural and computational mechanisms of how we arrive at affected beliefs.

[2] This study has been published as: Müller-Pinzler, L., **Czekalla, N.**, Mayer, A. V., Schröder, A., Stolz, D. S., Paulus, F. M., & Krach, S. (2022). Neurocomputational mechanisms of affected beliefs. *Communications Biology, 5*(1), 1241.
My contribution: designing the research, data acquisition, discussion of the data analyses and interpretation of the results, and review and editing of the manuscript.

3.2 Introduction

Self-efficacy can be defined as a person's subjective conviction that he/she can overcome challenging situations through his/her own actions (Bandura, 1977). To successfully perform goal-directed actions, humans must learn from incoming information, thereby forming beliefs about the world and about themselves enmeshed in this world. According to economic theory, learning should result in accurate beliefs that represent an internal model of the world that is suitable to inform decision making. Novel theoretical frameworks, among others by Bromberg-Martin and Sharot (Bromberg-Martin & Sharot, 2020), emphasize that besides the instrumentality (i.e. accuracy) of beliefs, they may also carry intrinsic value in and of themselves, thus shaping the learning process and how people ultimately arrive at their beliefs (Sharot & Garrett, 2016). In this regard, affective states, such as happiness about one's own good health prognosis, represent intrinsic values that individuals are inclined to optimize during belief formation (Bromberg-Martin & Sharot, 2020; Hughes & Zaki, 2015). To demonstrate this entanglement of affect and belief formation, we applied a learning task that induces affective reactions during the process of forming conceptually novel beliefs about one's abilities to master a task (Czekalla et al., 2021; Laura Müller-Pinzler et al., 2019). Specifically, we focused on the primary affective states elicited by self-efficacy beliefs – the self-conscious emotions of embarrassment and pride – and their impact on the beliefs. By exerting experimental control over failures and successes during the process of self-efficacy belief formation, we were able to assess how experienced affect relates to computational mechanisms of belief formation and the underlying activity of neural systems, linking neural and physiological mechanisms with shifts in preferences for information of positive or negative valence during learning.

Affective states are considered to guide cognitive processing, representing embodied and experiential information about the positive or negative value of what people encounter (Frijda, 1987; Storbeck & Clore, 2008). It is proposed that this internal affective information is integrated with external information to shape beliefs that rather than being objective, are motivated and biased by subjective feelings about the beliefs themselves, leading to a recursive influence of beliefs and affective states on each other (Bromberg-Martin & Sharot, 2020; Kunda, 1990; Loewenstein, 2006). Previous studies supported aspects of Bromberg-Martin and Sharot's framework (Bromberg-Martin & Sharot, 2020) by demonstrating that internal beliefs and external feedback can elicit emotions like happiness, pride, or embarrassment (Cecchi et al., 2022; Müller-Pinzler et al., 2015; Rutledge et al., 2016, 2014; Stolz et al., 2020; Vinckier et al., 2019, 2018). Affective states also have been shown to alter decision making (Charpentier, De Martino, et al., 2016; Charpentier, De Neve, et al., 2016; Stolz et al., 2020) and cognitive processes like situational judgments or learning styles (Storbeck & Clore, 2008). Social anxiety, low self-esteem, or depression, which are likely associated with more negative affective reactions to self-efficacy beliefs, have also been found to bias social learning (Koban et al., 2017; Korn et al., 2014; Müller-Pinzler et al., 2019; Will et al., 2020). These findings provide support for the overall rationale of the formation of affected beliefs, that is, the notion that

beliefs are fundamentally shaped by motivational biases as well as affective experiences during feedback processing. However, the question remains open of which neurophysiological mechanisms can explain how emotions elicited during learning are associated with biases in belief formation.

Neuroscientific studies provided initial evidence that common brain areas map the value of stimuli, actions, and their motivational relevance during social and non-social learning and decision making (Chib et al., 2009; Ruff & Fehr, 2014). Prediction errors, that is, the mismatch of prior expectation and a situation's outcome, are minimized by updating beliefs during learning. These are generally processed in the dopaminergically innervated ventral striatum, but also in the orbitofrontal cortex or the amygdala during learning (King-Casas et al., 2005; O'Doherty, 2004; Ruff & Fehr, 2014; Schultz et al., 1997). However, more recent findings suggest that there are distinct and unique neural computations which potentially reflect the impact of the prominent motivational and emotional processes during belief formation. For example, studies have shown that distinct value-related neural processes in subregions of the anterior cingulate cortex (ACC) are recruited depending on whether information about oneself or another agent is processed (Lockwood et al., 2016; Lockwood & Wittmann, 2018). Other findings revealed that activation in the ventral striatum was modulated when the social context changed from a private to a public situation, suggesting that the presence or absence of other people influenced the sensitivity to the reward value of certain decisions (Izuma et al., 2010). Biases specific to self-related learning, which are absent when one is learning about another person (Kuzmanovic et al., 2016; Müller-Pinzler et al., 2019), have been associated with differences in the tracking of negative prediction errors (Sharot et al., 2011). In this regard, the ventromedial prefrontal cortex (vmPFC) shows valence-specific encoding of self-related feedback, which has been shown to predict an optimism bias in belief updating (Kuzmanovic et al., 2016, 2018).

Affective states triggered after personal failures or successes are particularly important when people acquire novel self-concepts (Hopkins et al., 2021) and develop an initial understanding of themselves as being self-efficacious individuals in a novel task environment. Central to the entanglement of affect and such self-efficacy beliefs is the assumption that people are highly motivated to perform well and maintain or even construct a positively shaped self-image (Markus & Wurf, 1987; Sedikides & Gregg, 2008). Within this process, performance feedback elicits self-conscious emotions, such as pride in the case of success (Stolz et al., 2020; Tangney et al., 2007; Williams & DeSteno, 2008), but also embarrassment if one fails to achieve the expected outcome (Miller, 1996; Müller-Pinzler et al., 2015; Tangney et al., 2007). Self-conscious emotions differ from other emotional concepts as they essentially involve self-referential evaluations and the activation of self-concepts (Tangney et al., 2007). Thus, when it comes to emotional experiences in the context of a performance situation, pride or embarrassment are theoretically more valid constructs to capture differences in affective experiences than e.g., the basic emotion happiness. In the past, it has been demonstrated that these self-conscious emotions are not only a consequence of the situation but also directly affect

behavior. Pride experiences function as a motivator to persevere (Williams & DeSteno, 2008). In contrast, embarrassment experiences rather lead individuals to stop their current behavior, withdraw, and appease others (Apsler, 1975; Feinberg et al., 2012). For the process of belief formation, it has been argued that specifically the dorsomedial frontal cortex (dMFC), the ventral and dorsal anterior insula (vAI/ dAI), and the amygdala, brain areas involved in action monitoring as well as emotional processing, integrate affective states with outcome information (Koban & Pourtois, 2014). Therefore, the anterior insula (AI) has been regarded, among other brain regions, as an integrative hub for motivated cognition and emotional behavior (Koban & Pourtois, 2014; Wager & Barrett, 2017). Similarly, dopaminergic midbrain nuclei in the ventral tegmental area and substantia nigra (VTA/ SN) are associated with attention processes, and at the same time, with events (i.e. reward cues) that are of motivational relevance specifically during learning (Adcock et al., 2006; Schultz, 1998).

While current frameworks support the idea that intrinsic outcomes such as affective states may impact the formation of self-efficacy beliefs (Bromberg-Martin & Sharot, 2020; Hughes & Zaki, 2015), studies on this issue have not yet probed this framework as a whole. We aim to bridge this gap by showing how emotional states relate to biases in the formation of self-efficacy beliefs, and how they are associated with preferences for information of positive or negative valence. For this purpose, we tested the effects of individual differences in the affective reactions during learning. Using trial-by-trial updates of performance expectations in a conceptually novel task environment, we computed prediction error learning rates by fitting computational learning models revealing valence-specific learning biases. As predicted by current frameworks, individual differences in the experience of the emotions embarrassment and pride were distinctly related to biases in the formation of self-efficacy beliefs. Biased learning and affect were jointly related to the neural processing of valence-specific prediction errors in the AI, amygdala, VTA/ SN and mPFC as well as pupillary reactivity in favor of the preferentially used information to update the belief. Increases in valence-specific functional connectivity of the dAI with the amygdala, VTA/ SN and mPFC support the notion of an integrative mechanism of affective and motivational processes within the dAI (Kelly et al., 2012; Kurth et al., 2010). These findings provide insights into brain networks involved in computational biases associated with emotional experiences, and coherently support current theoretical frameworks integrating affective experiences in the process of belief formation.

3.3 Results

Measuring self-efficacy belief formation

In the present experiment, n = 39 subjects (26 females, aged 18-28 years; M = 22.3; SD = 2.65) completed the task in the MRI. Another n = 30 subjects (24 females, aged 18-32 years; M = 23.3; SD = 3.97) completed the task outside the MRI as a behavioral study. During the MRI scanning, eye-tracking data was additionally obtained in all but three subjects (see 3.5 Methods for more details). To examine the formation of self-efficacy beliefs we used the Learning Of Own Performance (LOOP) task (Czekalla et al., 2021;

Laura Müller-Pinzler et al., 2019).
In brief, in the LOOP task participants are asked to estimate specific characteristics of properties (e.g., heights of buildings, weights of animals, numbers of things, or distances between objects). By manipulating the performance feedback, participants are led to form novel beliefs on their own and the other person's cognitive estimation abilities. In the fMRI sample, participants perform the LOOP task in the MRI scanner while a confederate (presented as another participant) ostensibly performs the task simultaneously in an adjacent room. After each trial, participants receive a manipulated performance feedback for the last estimation (see Figure 3.1a). During the entire experiment, participants take turns in performing the estimation task themselves (Self condition) or observing the other participant performing the task (Other condition). Before each trial, participants are asked to rate either their own or the other person's expected performance for the upcoming trial, enabling us to examine the process of self- and other-related belief formation. The design of the LOOP task provides a High Ability and a Low Ability condition, resulting in overall four feedback conditions: Agent condition (Self vs. Other) x Ability condition (High Ability vs. Low Ability; see Figure 3.1b and 3.5 Methods for a detailed description of the task). In previous studies, we showed that over time, participants adjusted their expected performance ratings according to the feedback, allowing for an assessment of valence-specific self- and other-related learning processes (Czekalla et al., 2021; Laura Müller-Pinzler et al., 2019).

Selection of computational models for self-efficacy belief formation
Following a model-free behavioral analysis (see Supplementary Note 3.1), we modeled the participants' behavior by means of learning rates. Changes in expectations were modeled through updates from prediction errors (PEs) to test different learning rates for PEs with positive vs. negative valence and Self vs. Other (Supplementary Figure 3.1 and Supplementary Figure 3.2). In line with our previous studies, the winning model was a Valence Model, including separate learning rates for positive and negative PEs for Self vs. Other (Model 8; for a more detailed description of this model and the whole model space, see 3.5 Methods). This model received the highest sum PSIS-LOO score (approximate leave-one-out cross-validation (LOO) using Pareto smoothed importance sampling (PSIS, Vehtari et al., 2016) out of all models (for all PSIS-LOO scores see Supplementary Table 3.1). In addition, Bayesian model selection (BMS) resulted in a protected exceedance probability of $pxp > .999$ for this model and a Bayesian Omnibus Risk of $BOR < .001$. The expected model frequency was 46.53. Thus, the extended Valence Model was selected for all further analyses of learning parameters, allowing for a comparison of valence-specific learning rates. The time courses of performance expectation ratings predicted by our winning model successfully captured trial-by-trial changes in the actual expectations due to PE updates within each of the ability conditions at the individual subject level (R^2 = 0.46 ± 0.28 [*M* ± *SD*]), supporting the validity of the model in describing the subjects' learning behavior. In addition to revealing PE valence-specific learning, which could not be directly assessed via model-free behavioral analyses, posterior predictive checks also confirmed that the winning model captured the core effects in our model-free analysis (see

Supplementary Note 3.2, Figure 3.1c, Supplementary Table 3.2; for parameter correlations see Supplementary Table 3.3). Exploratory analyses with learning rates from Model 5 showed that our results were unaffected by the w parameter modulating learning from more extreme feedback in the winning Model 8 (see Supplementary Note 3.3).

Replication of the negativity bias for the formation of self-efficacy beliefs

Participants showed higher learning rates when forming self-efficacy beliefs than when forming beliefs about the other person's performance (main effect of Agent: $F_{(1,67)}$ = 5.77, p = .019, η^2 = 0.017, *partial* η^2 = 0.079). There was also a main effect of Prediction Error Sign ($F_{(1,67)}$ = 5.22, p = .025, η^2 = 0.011, *partial* η^2 = 0.072; categorical comparison of learning rates for positive vs. negative PEs) and a significant interaction of Agent x Prediction Error Sign ($F_{(1,67)}$ = 21.47, p < .001, η^2 = 0.040, *partial* η^2 = 0.243), which replicates earlier findings of a bias towards more negative updating during self-efficacy belief formation ($t_{(68)}$ = -3.53, p < .001, d = -0.425, $M\alpha_{Self/PE+}$ = 0.25, SD = 0.13; $M\alpha_{Self/PE-}$ = 0.35, SD = 0.20, Müller-Pinzler et al., 2019). Forming beliefs about another person's performance did not reveal a significant bias towards more negative updating ($t_{(68)}$ = 2.67, p = .009, d = 0.321; $M\alpha_{Other/}$ PE+= 0.27, SD = 0.16; $M\alpha_{Other/PE-}$ = 0.24, SD = 0.15; see Figure 3.1d). There was no significant main effect or interaction for Group (p > .097).

Associations of self-efficacy belief formation with affective experience

We hypothesized that self-efficacy belief formation is associated with affective experience. In line with Bromberg-Martin and Sharot (2020) we expected that individuals with more negative affective experience would update their self-efficacy beliefs in a more negative way. To quantify associations between learning behavior and affect, individual differences in the overall experience of embarrassment and pride during the task were used as between-subject measures. Embarrassment and pride ratings were only weakly correlated ($\rho_{(68)}$ = -.10, p = .436), indicating that the experience of embarrassment and pride during the task represent two rather independent affective components with respect to the self-related feedback (see Supplementary Table 3.4 for a more detailed correlation table). The bias in the formation of self-efficacy beliefs (Valence Learning Bias = ($\alpha_{Self/PE+}$ - $\alpha_{Self/PE-}$)/ ($\alpha_{Self/PE+}$ + $\alpha_{Self/PE-}$); Müller-Pinzler et al., 2019; Niv et al., 2012; Palminteri et al., 2017) was negatively linked to the reported experience of embarrassment during the task ($\rho_{(68)}$ = -.24, p = .043), that is, more negative updating behavior was associated with increased embarrassment ratings. In contrast, the Valence Learning Bias was positively linked to the emotion of pride ($\rho_{(68)}$ = .55, p < .001). A regression predicting the Valence Learning Bias with both affect ratings simultaneously revealed independent effects of pride (β = 0.56, $t_{(66)}$ = 5.81, p < .001) and embarrassment (β = -0.22, $t_{(66)}$ = -2.30, p = .025; R^2 = .41, $F_{(1,66)}$ = 22.90, p < .001, f^2 = 0.64). When controlling for differences in the feedback participants received before rating their affective experience, correlations between emotions and Valence Learning Bias do not significantly change and the overall pattern of associations remains consistent. This indicates that the experience of self-conscious

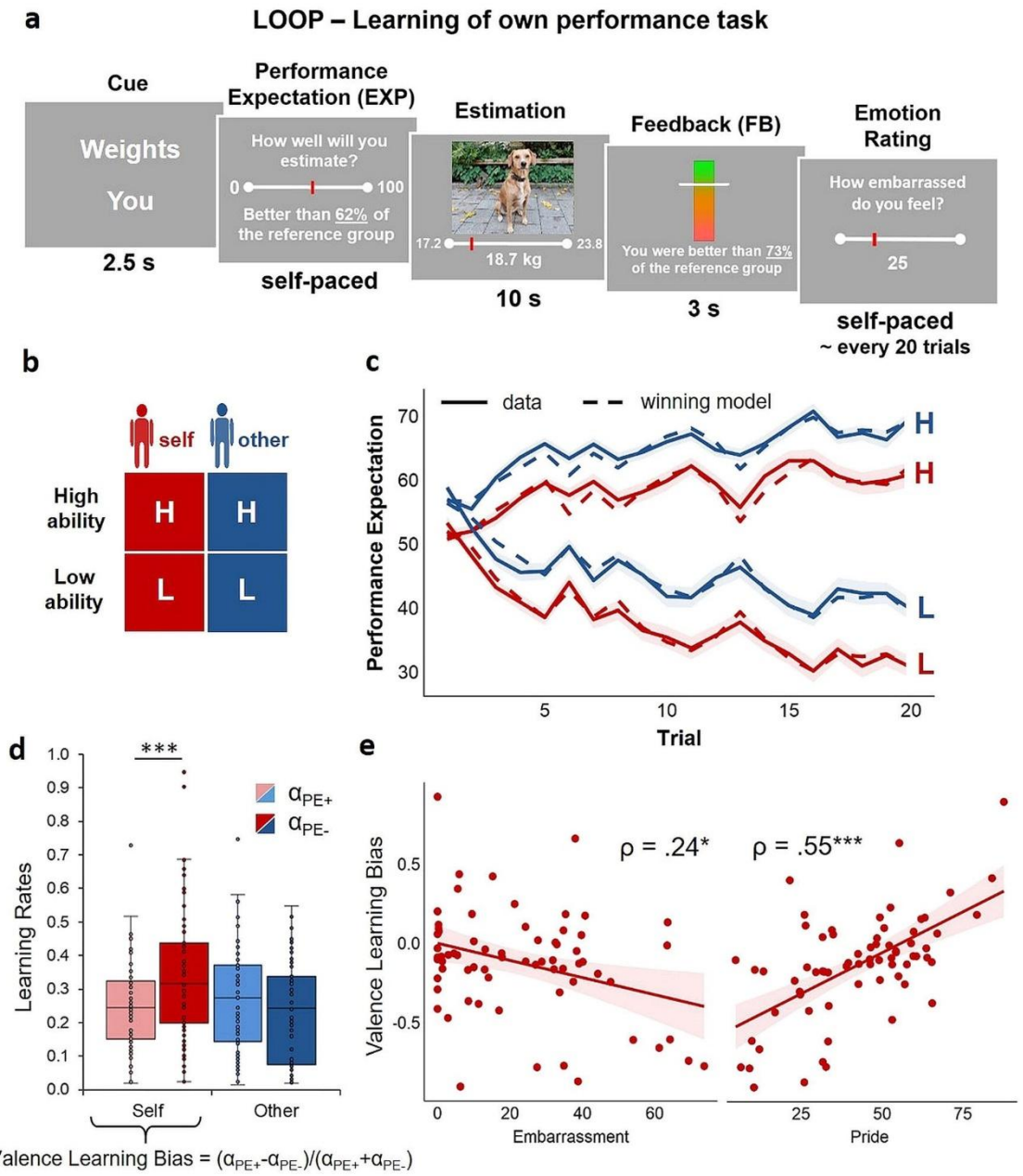


Figure 3.1. Trial sequence and timing, experimental conditions, modeling of learning behavior, learning rates and their association with self-conscious emotions. **a** Shows a stylized version of the estimation task. A cue at the beginning of each trial indicated the following estimation category and the agent whose turn it was. After providing their performance expectation ratings (EXP), participants were asked an estimation question, followed by the corresponding performance feedback. After approximately every 20 trials, participants were asked to rate their current emotional state (pride, embarrassment, happiness, stress/ arousal). **b** The LOOP task contained two experimental factors, Ability level (High Ability vs. Low Ability) and Agent (Self vs. Other), resulting in four feedback learning conditions that can be distinguished by different estimation question types (e.g. estimation of weights). **c** Predicted and actual performance expectation ratings across time. The behavioral data indicate that participants adapted their performance expectation ratings (solid lines) to the provided feedback, thus learning about the allegedly distinct performance levels. The winning valence-specific learning model captured the participants' behavior, as indicated by a close match of actual performance expectations with the predicted data (dashed lines). Shaded areas represent the standard errors for the actual performance expectations. **d** Learning rates derived from the winning Valence Model indicate that there was a bias towards increased updating in response to negative prediction errors (αPE-) in contrast to positive prediction errors (αPE+) for the formation of self-efficacy beliefs. Colored bars indicate the first and third quartile of the data, the line marks the median. Whiskers extend from the upper (lower) box borders to the largest (smallest) data point at most 1.5 times the interquartile range above (below) the respective border. Data with more extreme values than this are displayed as individual points; ***p < 0.001,

indicates a significant negativity bias during the formation of self-efficacy beliefs. **e** Correlation plots and Spearman correlations of self-related Valence Learning Bias and embarrassment as well as pride experience during the experiment. *$p < 0.05$, ***$p < 0.001$.

emotions during successful and unsuccessful performances was tied to the way in which people updated their self-efficacy beliefs (see Figure 3.1e). Furthermore, the way in which participants processed the performance feedback in order to update their self-efficacy beliefs was associated with their self-esteem. Specifically, participants with higher self-esteem showed more positive updating, $\rho_{(68)} = .33$, $p = .006$ (fMRI subsample: $\rho_{(38)} = .35$, $p = .030$), which strengthens the assumption that prior beliefs about the self have a direct impact on how individuals learn novel information about new abilities (Müller-Pinzler et al., 2019; Rouault et al., 2019).

Pupil dilation slopes are associated with surprise and valence of prediction errors, in line with a negative learning bias

Previous research has successfully linked changes in pupil diameter to surprise, PEs and learning (Koenig et al., 2018; Preuschoff et al., 2011) as well as emotional experiences and arousal (Bradley et al., 2008; Müller-Pinzler et al., 2015). Thus, we hypothesized that PE tracking is linked to changes in pupil diameter. To corroborate our assumption that changes in pupil diameter, as indicated by the slope of the change in pupil size during the processing of self-related feedback, reflect increased arousal or attention in association with greater PEs, we regressed trial-by-trial variability in the pupil slope on PE surprise (continuous effect of unsigned PEs) and PE valence (continuous effect of signed PEs; see Figure 3.2a; Rouhani & Niv, 2021). The linear mixed model revealed a significant positive effect for PE surprise ($\beta = 0.067$, $t_{(325.9)} = 2.16$, $p = .032$, *95% CI* = [0.006; 0.127]) and a significant negative effect for PE valence ($\beta = -0.113$, $t_{(30.7)} = -2.52$, $p = .017$, *95% CI* = [-0.200; -0.025]; see Supplementary Figure 3.3). First, we observed an effect of PE surprise, insofar as the more surprising the feedback was with respect to trial-by-trial prior expectations, the more the pupil dilated. Second, the results indicate that pupil dilation was greater with decreasing PE values, thus linking negative PEs, rather than positive PEs, to greater dilation (i.e. effect of PE valence). Potentially, these PE valence effects indicate increased arousal and attention towards more negative PEs, in line with the negativity bias that we found in learning rates.

Pupil dilation response to prediction error valence is associated with affect and learning bias

It has been suggested that pupil dilation reflects differences not only between stimuli but similarly between individual biases during decision making (see Figure 3.2b for examples of individual differences; de Gee et al., 2014). We thus expected individual differences in self-efficacy belief formation and affective experience to be associated with differences in pupil responses to PEs. To test this assumption, we introduced individual differences in learning and self-conscious emotions as between-subject covariates into the linear mixed models assessing trial-by-trial pupil slopes. These analyses demonstrated that individuals

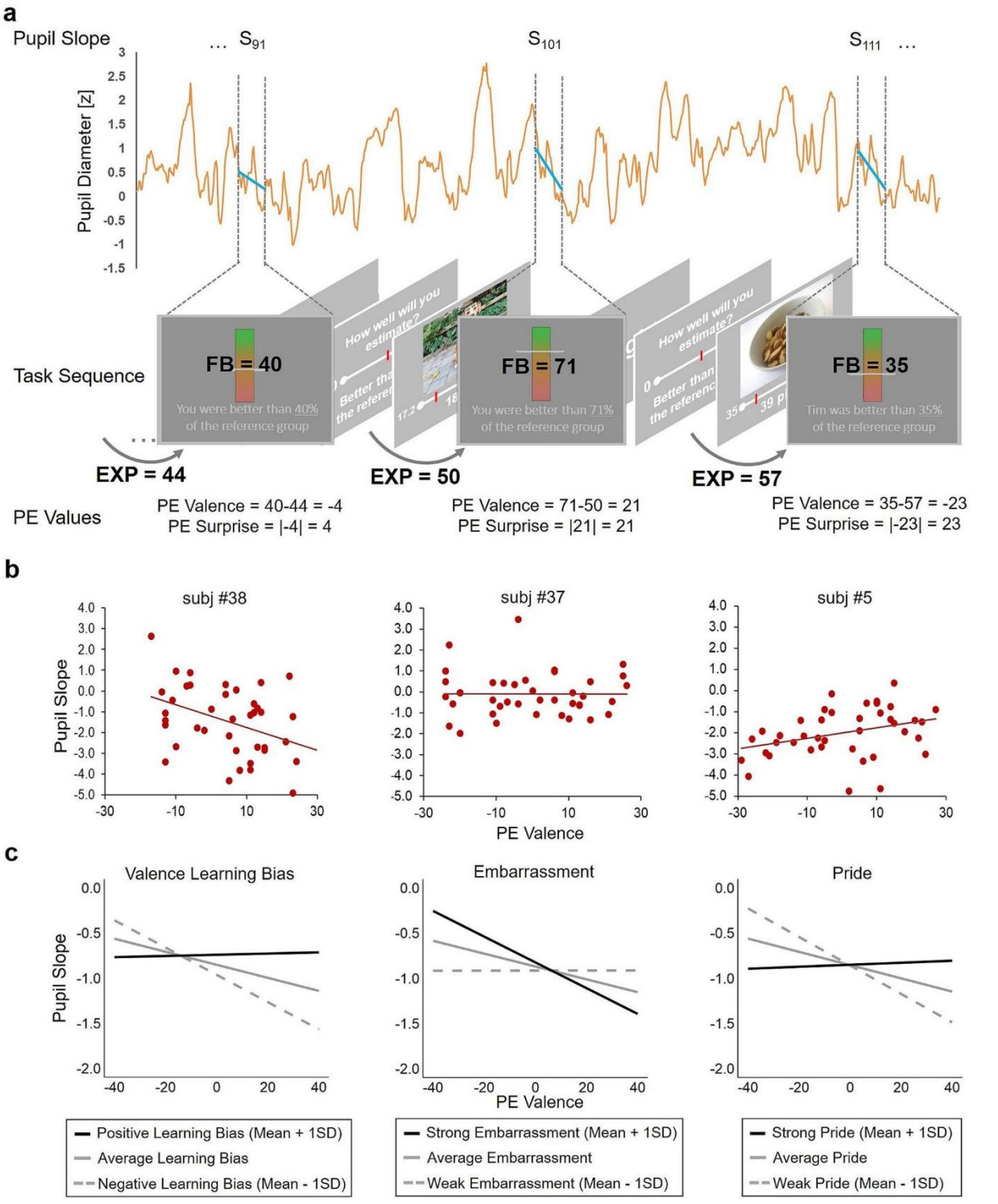


Figure 3.2. Association of pupil slopes with prediction error (PE) valence and individual pupil response differences explained by differences in Valence Learning Bias, embarrassment and pride experience. a Example of pupil diameter trace over three trials for one subject (orange line) and trial-specific fitted linear slopes (blue lines) for the feedback phase of each trial. PE values are calculated with the participant's current performance expectation (EXP) and the following feedback value (FB), and PE valence represents the signed PE while PE surprise represents the unsigned PE. b Three exemplary scatter plots show the association of pupil slopes with PE valence and illustrate the variance between subjects. Trend lines are fitted by linear regression. c Illustration of the impact of the three between-subject covariates, Valence Learning Bias (left), embarrassment (middle) and pride experience (right) explaining differences in the associations of PE valence and pupil slope. The plots show the data as predicted by the multi-level models for the mean covariate (grey line) and the mean covariate +/− 1 standard deviation (SD; black line and gray dashed line).

who experienced more embarrassment showed stronger pupil dilations scaling with more negative PEs, while pupil slopes did not correlate with PEs in individuals with lower embarrassment (significant interaction of embarrassment and PE valence: (β = -0.0004, $t_{(32.5)}$ = -2.59, p = .015, *95% CI* = [-0.0006; -0.0001]; no main effect for embarrassment: β = 0.002, $t_{(34.2)}$ = 0.42, p = .679, *95% CI* = [-0.009; 0.014]; see Figure 3.2c). These effects were reversed when pride ratings, instead of embarrassment ratings, were included in the model (interaction pride and PE valence (β = 0.0005, $t_{(34)}$ = 3.18, p = .003, *95% CI* = [0.0002; 0.0007]; main effect of pride: β = -0.00006, $t_{(34.1)}$ = -0.01, p = .991, *95% CI* = [-0.01132; 0.01120]). The Valence Learning Bias modulated the relationship between PE valence and pupil slopes in the same way (interaction Valence Learning Bias and PE valence (β = 0.38, $t_{(31.8)}$ = 3.02, p = .005, 95% CI = [0.13; 0.62]; main effect of Valence Learning Bias: β = 0.33, $t_{(50.9)}$ = 1.04, p = .308, 95% CI = [-0.30; 0.97]), indicating that participants with a more negative Valence Learning Bias showed a negative correlation of pupil dilation and PEs, whereas participants with no bias or a positive bias showed less differentiation in pupil dilation in response to the valence of the PE.

Common neural activations associated with PE surprise and distinct activations for PE valence

On the level of the brain, we assessed the association of PE tracking with neural activity and tested whether there is a specific response pattern with respect to self- and other-related belief formation. To do so, we computed the effects of continuous trial-by-trial PE surprise and PE valence as parametric weights to assess neural aspects of learning more specifically (see Figure 3.3a). Increased PE surprise was associated with greater activation of the mPFC for Self and Other as well as clusters in the left insula/ temporal pole/ frontal orbital gyrus (bilaterally for Other; see Figure 3.3c and Supplementary Table 3.5). There was no significant difference between Self and Other (p < .001 uncorrected), indicating that there is no evidence for distinct neural processes of error tracking between agents.

The assessment of PE valence revealed a distinct pattern for self- and other-related belief formation: Self-related PE valence was positively associated with increased activation of the NAcc/ VS, mPFC, bilateral angular gyrus/ superior parietal lobule/ lateral occipital gyrus and precentral gyrus, showing stronger activation scaling with more positive PEs (Figure 3.3b and Supplementary Table 3.6). There was no effect for other-related PE valence, and a direct comparison of self- vs. other-related PE valence effects revealed stronger associations in the NAcc/VS for Self (right: x, y, z: 12, 17, -4, $t_{(38)}$ = 5.23; k = 2; left: x, y, z: -9, 26, −1, $t_{(38)}$ = 5.77, k = 19). This supports the assumption that the valence of the feedback has a specific value when feedback refers to the self as compared to another person. Although behavioral data and learning rates clearly emphasize the greater importance of negative over positive PEs, there were no significant negative associations with PE valence in the neural data (p < .001 uncorrected). Additional analyses assessing differences between the feedback conditions for Agent and Prediction

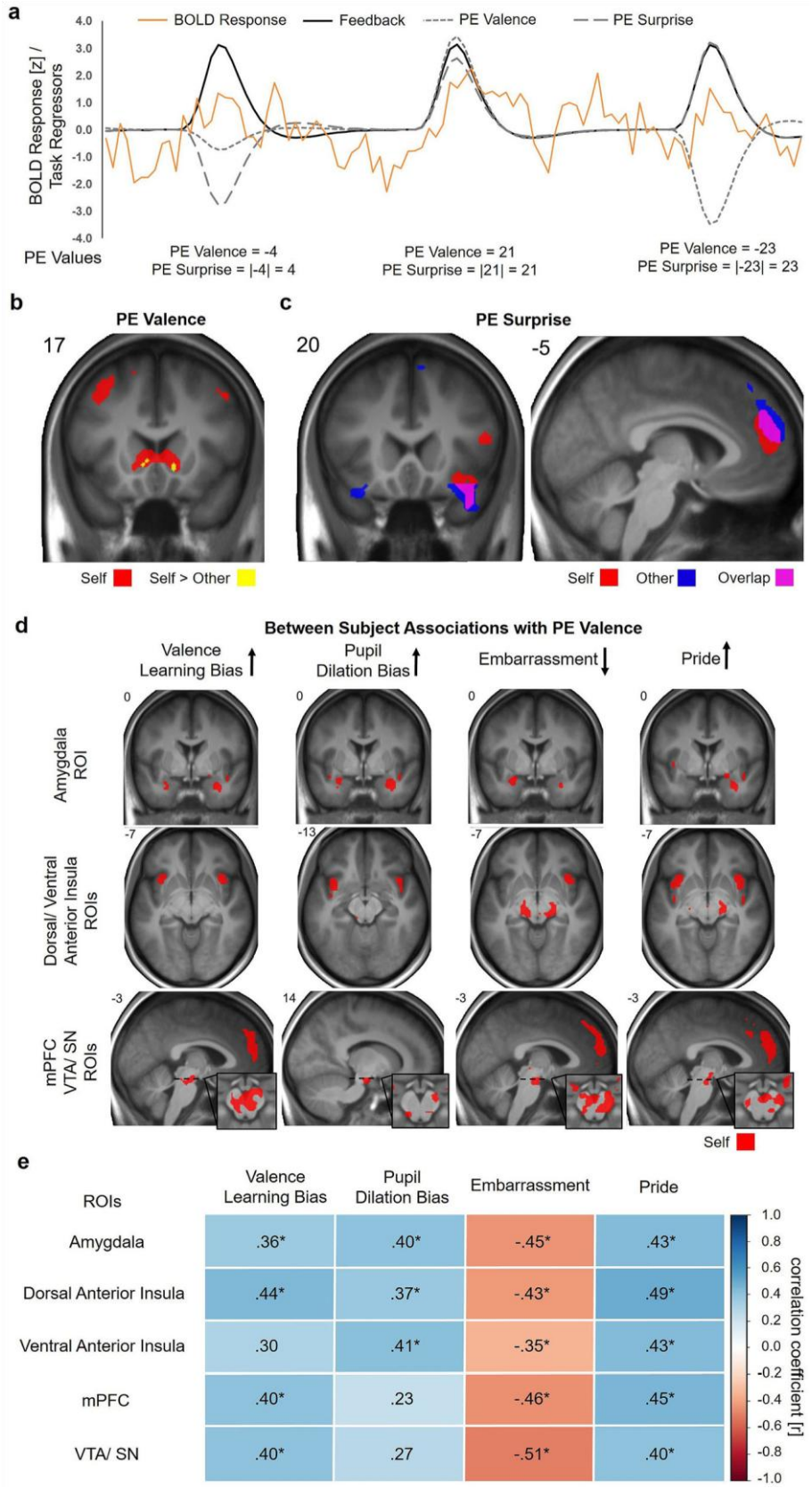


Figure 3.3. Common neural activations associated with prediction error (PE) surprise, distinct activations for self-related PE valence and individual response differences to PE valence explained by differences in Valence Learning Bias, embarrassment and pride experience, and pupil dilation. **a** Exemplary BOLD response over three trials for one subject (orange line) and regressors for the feedback phase of each trial (black line; the originally separate regressors for self- and other-related feedback are combined here for display purposes). PE valence (small dashed) and PE surprise (large dashed) are added as parametric modulators in addition to the feedback regressors. PE values are

calculated as shown in Figure 3.2. **b** PE valence was associated with increased activation of the NAcc/VS, mPFC, bilateral angular gyrus/ superior parietal lobule/ lateral occipital gyrus and precentral gyrus when participants formed self-efficacy beliefs (Self). **c** PE surprise was associated with activation of the mPFC and the bilateral insula/ temporal pole/ frontal orbital gyrus during the formation of self- and other related beliefs (uncorrected $p < 0.001$ for display purposes; see Supplementary Table 3.5 for FWE corrected statistics). **d** Neural tracking of PE valence during the formation of self-efficacy beliefs was modulated by between-subject variables. Black arrows indicate the direction in which the covariates are coded in the analyses. Clusters refer to $p < 0.005$, uncorrected for display purposes; see Supplementary Data 3.2 for FWE corrected statistics. **e** Pearson correlations for parameter estimates derived from the whole areas of our predefined ROIs with the Valence Learning Bias, Pupil Dilation Bias, embarrassment and pride are color-coded. *$p < 0.05$, FDR corrected.

Error Sign are presented in the Supplementary Information (Supplementary Note 3.4, Supplementary Figure 3.4, Supplementary Data 3.1, and Supplementary Table 3.7).

Neural activity in response to self-related PE valence is associated with affect, learning bias, and pupil dilation

To assess how biases in learning as well as affective experience and pupil dilation were associated with valence-specific PE processing on the single trial level, multiple general linear models (GLMs) were performed. These included the Valence Learning Bias, self-conscious emotions, and a score representing a valence bias for pupil dilation responses to positive vs. negative PEs (Pupil Dilation Bias = $PupilSlope_{Self/PE+}$ - $PupilSlope_{Self/PE-}$) as between-subject covariates for PE valence tracking. Analyses within our predefined regions of interest (ROIs) revealed that the more negative the Valence Learning Bias was, the more neural activity increased with more negative PEs in the bilateral dAI, vAI, amygdala, mPFC, and VTA/ SN (all results are $p < .05$ family-wise error (FWE) corrected at peak level within ROIs; see Figure 3.3d and Figure 3.3e, Supplementary Data 3.2). In other words, the more positive participants learned about themselves (i.e., more positive Valence Learning Bias), the more neural activity increased with more positive PEs in these regions (see Figure 3.3d, Figure 3.3e, and Supplementary Figure 3.5). Overall, higher experience of embarrassment showed similar associations with stronger activity with more negative PEs in the right dAI, bilateral amygdala, and VTA/ SN. Trend effects for embarrassment were found in the left dAI, bilateral vAI, and mPFC. In line with this, lower experience of pride showed the same association in the dAI and vAI, amygdala, VTA/ SN and mPFC. Additional analyses revealed that effects for embarrassment and pride were mainly independent (see Supplementary Note 3.5). Similarly, the more negative the Pupil Dilation Bias was, the stronger the activation of the dAI and vAI, amygdala and VTA/ SN with more negative PEs. Thus, the greater the response of this neural system for more negative PEs, the greater was the preference for negative information during learning as well as the negativity of the affective experience. This gained multi-modal support by similar associations of the Valence Learning Bias and affect with the pupil dilation response, which reflects the activity of this underlying neural system. In contrast, participants who showed a greater response of this neural system to positive PEs also had a preference for positive information during learning and reported more positive affect.

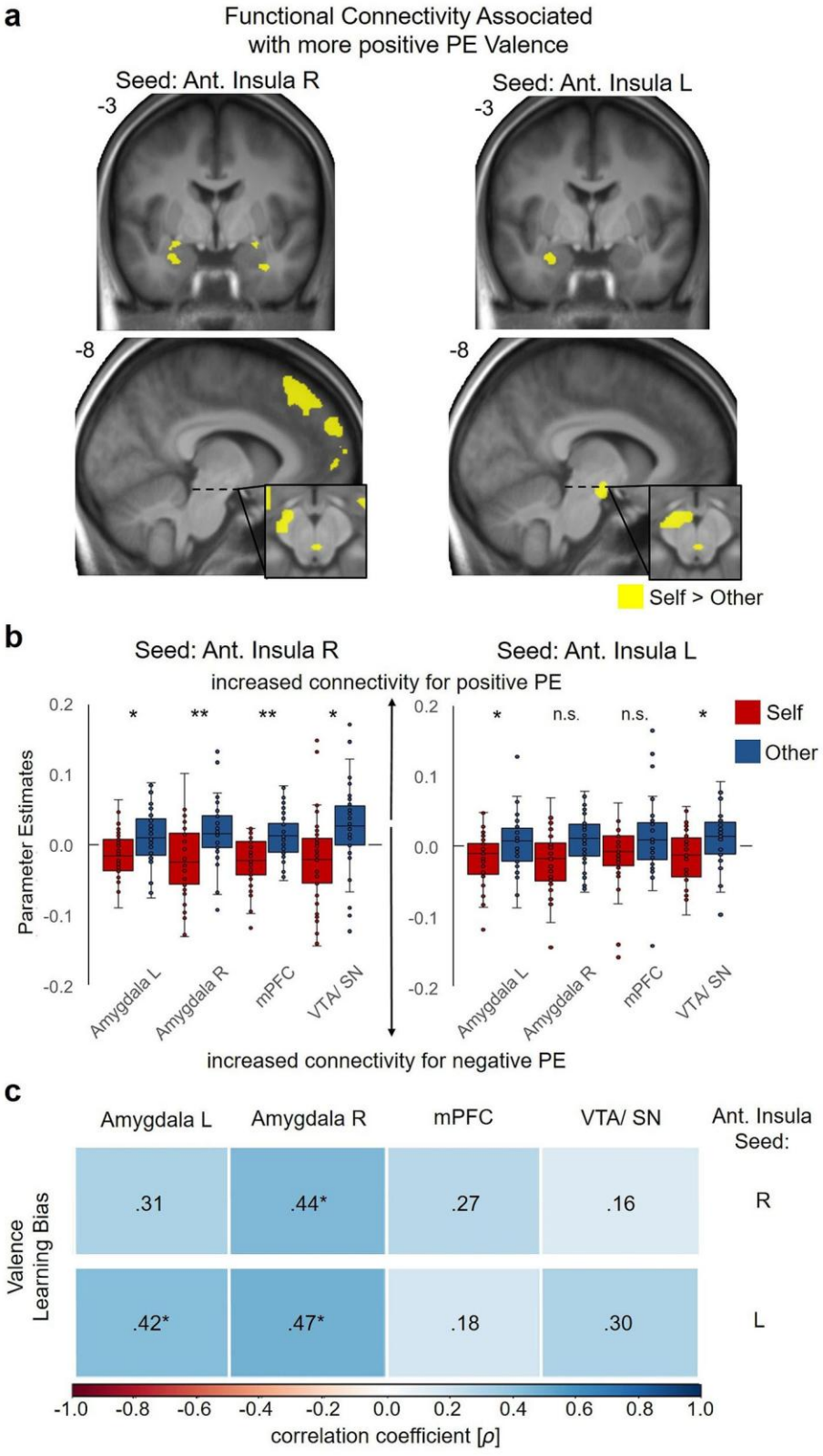


Figure 3.4. Differences in functional connectivity of the dorsal anterior insula during prediction error (PE) valence tracking during the formation of selfand other-related beliefs and associations with the Valence Learning Bias. a Increased functional connectivity of the dorsal anterior insula for the negative effect of PE valence in the predefined ROIs (amygdala, mPFC, VTA/ SN; $p < 0.005$ uncorrected for display purposes; contrast Self vs. Other). b Functional connectivity dynamics of the dorsal anterior insula plotted separately for the formation of self- and other-related beliefs. For display purposes, parameter estimates are plotted separately for Self and Other and refer to the peak voxels of the contrast Self vs. Other that are reported in Supplementary Table 3.8. Colored bars indicate the first and third quartile of the data, the line marks the median. Whiskers extend from the upper (lower) box borders to the largest

(smallest) data point at most 1.5 times the interquartile range above (below) the respective border. Data with more extreme values than this are displayed as individual points; *p < 0.05, **p < 0.01. c Spearman correlations of the Valence Learning Bias with the functional connectivity dynamics between the dorsal anterior insula (seed region reported on the right side) and the amygdala, mPFC and VTA/ SN associated with PE valence for self- vs. other-related learning are color-coded. *p < 0.05, FDR corrected.

Functional connectivity of the dorsal anterior insula depends on prediction error valence in line with the negativity bias

Due to the dense anatomical and functional connections between the dAI and (para)limbic as well as frontal brain regions (Kelly et al., 2012; Kurth et al., 2010) we tested whether the dAI connectivity was increased with more negative PEs. To do so, we assessed functional connectivity dynamics of the left and right dAI, as these were activated during feedback processing for self- and other-related feedback, independent of Agent and PE valence. Using psychophysiological interaction (PPI) analyses we calculated the interaction of the continuous PE valence and the time series extracted from the left and right dAI seed regions separately for Self and Other on the first level. The two agents were contrasted against each other on the second-level GLM, as we were specifically interested in connectivity dynamics that might reflect the differential learning from negative PEs when processing self-relevant information. Contrasting the PPI effects for PE valence between Self and Other demonstrated that during the formation of self-efficacy beliefs, functional connectivity dynamics of the right dAI with the bilateral amygdala, mPFC and VTA/ SN ($p < .05$, FWE corrected at peak level within ROIs) more strongly aligned with the negativity of the PEs. The left dAI showed a weaker but similar spatial distribution, with significant differences between self- and other-related PE valence for the left amygdala and VTA/ SN ($p <. 05$, FWE-corrected, see Figure 3.4a/ b and Supplementary Table 3.8). Thus, those brain regions that preferentially tracked PEs of negative valence in individuals with increased negative affect and learning biases also showed connectivity dynamics with the dAI in a similar direction during self-efficacy belief formation. Individuals who showed more pronounced differences in functional connectivity, that is, stronger functional connectivity for negative PEs during Self>Other, also showed a more negative Valence Learning Bias, although this pattern was not fully consistent across all ROIs (see Figure 3.4c).

3.4 Discussion

Belief formation is essentially biased, and various studies have shown how it is shaped by motivations (Elder et al., 2022; Sedikides & Gregg, 2008; Sedikides & Hepper, 2009; Sharot & Garrett, 2016). Here, we extend these findings and show that the affect, which people experience during learning, also is linked to belief formation and its underlying neural processes. Our computational modeling results imply that biases in the formation of self-efficacy beliefs in mastering a conceptually novel task are associated with the experience of the self-conscious emotions of embarrassment and pride. Critically, on the level of neural systems, the valence of prediction errors (PEs) is associated with biases in self-efficacy belief formation and the negativity of the affective experience. Individual differences in the response preference for negative PEs, as indicated by the pupil dilation

response and activation of the AI, amygdala, mPFC, and VTA/ SN, are associated with a more negative learning bias and negative affective experience, hinting at a neurobiological system that integrates affect during learning.

The novel framework on the value of beliefs proposed by Bromberg-Martin and Sharot (Bromberg-Martin & Sharot, 2020) nicely details how beliefs elicit emotions, while at the same time, these emotions shape how beliefs are updated in a reciprocal relationship. Based on this framework as well as previous research on self-conscious emotions, a negative belief about one's abilities, i.e. a negative self-efficacy belief, should elicit stronger embarrassment after failures and reduced pride after successes (Müller-Pinzler et al., 2015; Tangney et al., 2007). According to the present data, the association of the learning bias with affect supports this notion, as individuals who experienced more negative (embarrassment) and less positive affect (pride) when receiving self-related feedback were also inclined to update their self-efficacy beliefs in a more negative way. At the same time, negative emotions guide the information processing at various stages, including perception, attention, and decision-making, as discussed in the context of motivated cognition (Hughes & Zaki, 2015). This reciprocal relationship finally results in a biased formation of self-efficacy beliefs and in self-efficacy beliefs that are both drivers of affect and influenced by emotional responses to incoming information. Here, embarrassment is a particularly relevant example illustrating this recursive relationship: The fear of failure, as is often discussed in the context of social anxiety (disorder) (Koban et al., 2017; Morrison & Heimberg, 2013; Müller-Pinzler et al., 2015; Müller-Pinzler et al., 2019), leads to shifts in expectations and attention (threat monitoring) towards negative information. At best, this results in reparative behavior and performance improvement (Darby & Harris, 2010; Keltner & Potegal, 1997), and at worst, it leads to a vicious cycle of fear and pathologically increasing negative beliefs about the self (Heimberg et al., 2010). This is reflected in the present findings, when individuals who experienced more intense embarrassment ended up with lower self-efficacy beliefs.

Emotions shape learning processes in different ways. First, emotions can influence how information is processed in the brain by adaptively shifting attention towards salient aspects of the situation (Christianson, 2014; Kaspar & König, 2012). Second, emotions entail arousal, which intensifies internal rehearsal and evaluations, leading to increased learning (Christianson, 2014; Frijda, 1987; Storbeck & Clore, 2008), although these processes often interact and are intricately related (Hughes & Zaki, 2015). The increased pupil dilation in response to negative PEs in our study is in line with both increased salience of and attentional shifts towards negative PEs (Koenig et al., 2018; Preuschoff et al., 2011) or increased arousal elicited by negative PEs (Bradley et al., 2008; Müller-Pinzler et al., 2015). In this regard, we believe that the stronger impact of positive or negative information on pupil responses and brain reactivity maps arousal and affect according to the valence of individual learning biases and affective experiences.

On the neural level, the anterior insula (AI), specifically, has been suggested to function as an integrative hub for motivated cognition and emotional behavior (Koban & Pourtois, 2014; Wager & Barrett, 2017). While ventral aspects of the AI are associated with affective processing, emotions, and physiological arousal (Craig, 2003; Kelly et al., 2012; Lindquist et al., 2012; Phan et al., 2002; Wager & Barrett, 2017), dorsal aspects of the AI are

strongly associated with the detection of salient events, allocation of attentional resources, executive working memory (Menon & Uddin, 2010; Touroutoglou et al., 2012) and also surprise PEs and uncertainty during learning (Loued-Khenissi et al., 2020; Rutledge et al., 2010; Ullsperger et al., 2010). These findings suggest that the functions of the AI provide a physiological basis for how emotions are translated into biased, motivated, or affected beliefs (Koban & Pourtois, 2014; Wager & Barrett, 2017). A similar role, as a link for the attention-emotion interaction, has also been suggested for the amygdala (Kaspar & König, 2012; Koban & Pourtois, 2014), which showed similar responses in our task. The functional connectivity dynamics of the dAI, matching the modeled learning rates with a stronger impact of self-related negative PEs, underline the insula's role as an integrative hub (Mesulam & Mufson, 1982) that receives and forwards signals affecting information processing in other brain regions (Kelly et al., 2012; Kurth et al., 2010).

Tracking of PEs in the dopaminergically innervated VTA/ SN is influenced by motivational factors during learning (Adcock et al., 2006). The subjective value of self-related information varies strongly between subjects, as indicated by response patterns of the VTA/ SN to gains or losses (Charpentier et al., 2018). In this line, we believe that the present results reflect individual response tendencies at a very basic level of PE tracking. On higher layers of the computational hierarchy, regions in the ACC and mPFC are also associated with PE tracking and value representation (Hare et al., 2008; Lockwood & Wittmann, 2018; Wallis & Kennerley, 2010) and have been previously associated with biases in learning (Korn et al., 2012; Kuzmanovic et al., 2016, 2018). Affect and arousal could therefore bias learning on various stages of the computational hierarchy of PE processing, from more basic dopaminergic midbrain responses to more abstract value representations in the neocortex (Diaconescu et al., 2017). While the directionality of the effects remains to be determined, the dynamics in the functional connectivity of the dAI suggest a modulatory role in this process. Here, information is forwarded to and/ or integrated from the VTA/ SN and mPFC, the same regions whose response to the valence of PEs was also modulated by differences in learning bias and affective experience. This strengthens the idea that the AI may play a role in shifting responses to negative or positive information in other brain regions (e.g. by shifting attention and by affective tagging) or that it already may receive stronger signals in response to PEs of negative or positive valence from midbrain regions and the mPFC. The tracking of the absolute error, PE surprise (Rouhani & Niv, 2021), independently of the agent, suggests that there is a common and valence-independent coding of surprise in the insula and the mPFC that could be sufficient to complete the learning task per se. As our results indicate, however, the valence of the PE is relevant for understanding the trajectories of how individuals form self-efficacy beliefs. This is implicated by a valence-dependent, additive shift in the error-related BOLD response of these regions that corresponds to individual differences in the learning bias and affective experience. As a result, besides the main effect of surprise on the BOLD response of the AI and the mPFC, individuals who form more negatively biased self-efficacy beliefs and experience more negative affect, also have greater error-related responses in the case of negative PEs in contrast to positive PEs. This congruency in the modulation of the U-shaped surprise function hints at a neurocomputational mechanism of how affect may shape the formation of beliefs, as proposed previously (Bromberg-

Martin & Sharot, 2020). Some of the key findings of the present study emerged at the level of individual differences. We observed a wide inter-individual variance in the affective experience during the task and in the learning bias, that is, the type of information participants preferentially used to update self-efficacy beliefs. While, on average, we found a negativity bias during the formation of self-efficacy beliefs, just under a third of the participants still showed a positive learning bias, underlining the importance of individual factors and the meaningfulness of variability. Studies suggest not only that biases in belief formation differ between tasks (Ertac, 2011; Müller-Pinzler et al., 2019; Sharot & Garrett, 2016) but also that they depend on contextual factors like stress (Czekalla et al., 2021; Garrett et al., 2018). An individual's ability to adjust his/her current information processing strategy to the context might be adaptive (Bromberg-Martin & Sharot, 2020): for example, adaptation to an increased relevance of negative or threat-related information during stress (Garrett et al., 2018) or coping with a negative self-concept following social stress by means of more self-beneficial belief updating (Czekalla et al., 2021). It might also be adaptive for people who fear negative feedback to pay more attention to failure-related information in order to learn and circumvent potential future failures (Sedikides & Hepper, 2009). However, it is not always straightforward to determine under which conditions a strategy is adaptive or whether the affective experience can ameliorate the individual's well-being. A maladaptive consequence of biased self-efficacy beliefs becomes apparent in psychiatric disorders such as depression and social anxiety, in which amplified negative updating can lead to persistently distorted self-views and overly negative beliefs about one's own capabilities in everyday life (Alden et al., 2008; Amir et al., 2012; Koban et al., 2017; Korn et al., 2014; Taylor & Brown, 1988).

Emotions experienced during learning are linked to computational mechanisms and manifest in distributed neural activity during belief formation. In particular, neural activity of the AI, amygdala, VTA/SN, and mPFC and pupil responses map the valence of PEs in correspondence with the experienced affect and the learning bias that people show during belief formation. The more negative balancing in the functional connectivity dynamics of the dAI during the processing of self-related PEs within this network outline a scaffold for neural and computational mechanisms integrating affect during belief formation. The results of our empirical implementation of the framework on the value of beliefs (Bromberg-Martin & Sharot, 2020) have broader implications concerning any context that provides personal evaluations based on behavioral performance. Here, the focus on the affective experience during learning provides a deeper understanding of how feedback manifests in self-efficacy beliefs, which may in turn have a relevant impact on developmental processes and future behavior.

3.5 Methods

Participants

The study was approved by the ethics committee of the University of Lübeck (AZ 18-066), was conducted in compliance with the ethical guidelines of the American Psychological Association (APA), and all subjects gave written informed consent. Participants were recruited at the University Campus of Lübeck, were fluent in German, and had normal or corrected-to-normal vision. In the MRI 39 participants (26 females, aged 18-28 years; *M*

= 22.3; *SD* = 2.65) completed the study. We initially recruited 48 participants, but had to exclude six participants who did not believe the cover story of the task and three participants who did not attentively complete the task until the end (e.g. participants reported that they were too tired or the ratings indicated that they stopped responding to the estimation task). During the MRI scanning, eye-tracking data was additionally obtained and could be analyzed in all but three subjects who had insufficient data quality (resulting in *n* = 36 for pupil data analyses). We recruited an additional 30 participants (24 females, aged 18-32 years; *M* = 23.3; *SD* = 3.97), who completed the study as a behavioral study outside the MRI to increase the sample size for computational modeling results (resulting in an overall *N* = 69 for behavioral data analyses). For more details on the sample characteristics, see Supplementary Table 3.9.

Learning of own performance task

The learning of own performance (LOOP) task enables participants to incrementally learn about their own or another person's alleged ability in estimating properties. The task was previously introduced and validated in a set of behavioral studies (Müller-Pinzler et al., 2019). For the LOOP task, all participants were invited to take part in an experiment on cognitive estimation together with a confederate, who was allegedly another participant. In contrast to the fMRI study, for the behavioral study, two participants were invited and tested together instead of introducing a confederate. Participants were informed that they would take turns with the other participant/ confederate, either performing the task themselves (Self) or observing the other person performing (Other). Participants were asked to estimate different properties (e.g. the height of houses or the weight of animals). On a trial-by-trial basis, participants received manipulated performance feedback in two distinct estimation categories for their own estimation performance and for the other person's estimation performance. Unbeknownst to the participant, one of the two categories was arbitrarily paired with rather positive feedback while the other was paired with rather negative feedback (e.g. height of houses = High Ability condition and weight of animals = Low Ability condition or vice versa; estimation categories were counterbalanced between Ability conditions and Agent [Self vs. Other] conditions). This resulted in four feedback conditions with 20 trials each (Agent condition [Self vs. Other] x Ability condition [High Ability vs. Low Ability]). Trials of all conditions were intermixed in a fixed order with a maximum of two consecutive trials of the same condition. Performance feedback was provided after every estimation trial, indicating the participant's own or the other person's current estimation accuracy as percentiles compared to an alleged reference group of 350 university students who, according to the cover story, had been tested beforehand (e.g. "You are better than 94% of the reference participants."; see Figure 2.4a). The feedback was defined by a sequence of fixed PEs with respect to the participants' current belief about their abilities. The current belief was calculated as the average of the last five performance expectation ratings per category, which started at 50% before participants actually rated their performance expectation. This procedure led to varying feedback sequences between participants but kept PEs mostly independent of the participants' performance expectations and ensured a relatively equal distribution of

negative and positive PEs across conditions (Self: mean positive PE = 13.6, SD = 1.8 (mean frequency = 20.3); mean negative PE = -12.6, SD = 1.4 (mean frequency = 19.7); Other: mean positive PE = 13.0, SD = 1.3 (mean frequency = 19); mean negative PE = -13.1, SD = 1.1 (mean frequency = 21)). At the beginning of each trial, a cue was presented indicating the estimation category (e.g. height) and the agent whose turn it was (e.g. you or Tim). Afterwards participants were asked to state their expected performance for this trial on a scale with the same percentiles used for feedback. In order to increase motivation and encourage honest response behavior, participants were informed as part of the cover story that accurate expected performance ratings would be rewarded with up to 6 cents per trial, that is, the better the match between their expected performance rating and their actual feedback percentile, the more money they would receive. Following each performance expectation rating, the estimation question was presented for 10 s. During the estimation period, continuous response scales below the pictures determined a range of plausible answers for each question. Participants indicated their responses by navigating a pointer on the response scale with an MRI-compatible computer mouse. Subsequently, feedback was presented for 3 seconds (see Figure 2.4a). Jittered inter-stimulus-intervals were presented following the cue (mean: 4 * TR (0.992 s), range: 2-6 * TR), estimation (mean: 4.5 * TR, range: 2.5 – 6.5 * TR) and feedback phase (mean: 6 * TR, range: 4-8 * TR) for the fMRI task with jitters distributed in a uniform distribution with steps of 0.5 * TR. All stimuli were presented using MATLAB Release 2015b (The MathWorks, Inc.) and the Psychophysics Toolbox (Brainard, 1997). The fMRI task was completed in two separate 20-min sessions with a short break in between.

Before starting the experiment, all participants answered several questions about their self-efficacy beliefs and completed a self-esteem personality questionnaire (Self-Description Questionnaire-III, SDQ-III; Marsh & O'Neill, 1984). During the LOOP task, participants were also asked to rate their current levels of embarrassment, pride, happiness and stress/ arousal on a continuous scale ranging from not at all (coded as 0) to very strong (coded as 100). During the whole task four emotion rating phases, including all four emotions, were presented, each following a trial of one of the four experimental conditions (e.g. Self - High Ability). The two emotion rating phases following self-related feedback were averaged to obtain a rating for the experience of self-conscious affect (embarrassment and pride) during the formation of self-efficacy beliefs. Following the task, participants completed an interview including ratings about self-efficacy beliefs, were debriefed about the cover story, and reimbursed for their time before leaving. The whole procedure took approximately 2 h.

Statistics and reproducibility

Behavioral data analysis and modeling. To illustrate effects in our behavioral data, a model-free analysis was performed on the participants' expected performance ratings for each trial. We conducted a linear mixed model (LMM) fitted with restricted maximum likelihood (REML) including the Ability condition (High Ability vs. Low Ability) x Agent condition (Self vs. Other) as factorial and Trial (20 Trials) as continuous predictors.

Intercept, Ability condition, Agent condition, and Trial were modeled as fixed and random effects (see Supplementary Note 3.1 for results).

Following model free analyses, dynamic changes in self-efficacy beliefs, that is, performance expectation ratings, were then modeled using PE delta-rule update equations (adapted Rescorla-Wagner model; Rescorla & Wagner, 1972). For the learning models the following PE delta-rule update equation was used (EXP = Performance expectation rating, FB = feedback, PE = prediction error, α = learning rate):

[1] $EXP_{t+1} = EXP_t + \alpha\ PE_t$; while $PE_t = FB_t - EXP_t$

The model space contained three main models, which varied with regard to their assumptions about biased updating behavior when forming self-efficacy beliefs (see Supplementary Figure 3.1). The simplest learning model used one single learning rate for all conditions for each participant, thus not assuming any learning biases (Unity Model). The second model, the Valence Model, included separate learning rates for positive PEs (α_{PE+}) vs. negative PEs (α_{PE-}) across both ability conditions, thus suggesting that the valence (positive vs. negative) of the PE biases self-efficacy belief formation. The third model, the Ability Model, contained a separate learning rate for each of the ability conditions, indicating context-specific learning. In addition, learning rates were either estimated separately for Self vs. Other or across Agent conditions. The Valence Model with separate learning rates for Self vs. Other (Model 5), which was the winning model in our previous studies (Czekalla et al., 2021; Müller-Pinzler et al., 2019), was further extended by adding a weighting factor that reduced learning rates towards the ends of the feedback scale (percentiles close to 0 % or 100 %), under the assumption that participants would perceive extreme feedback values to be less likely than more average feedback (Kube et al., 2021). In the first of these models (Model 7), a linear decrease of the learning rates was assumed, beginning at 50 % and ending at 0 % and 100 %. A weighting factor w was fitted for each participant, defining how strongly the linear decrease was present for each individual. Since many of the variables people encounter in everyday life (e.g., many test results) approximately follow a normal distribution with extreme values being less likely, for the second model of this kind (Model 8), we assigned the relative probability density of the normal distribution to each feedback percentile value. Again, a weighting factor w was fitted for each individual, indicating how strongly the relative probability density reduced the learning rates for feedback further away from the mean. The initial beliefs about the own and the other participant's performance (EXP_1) were estimated as free parameters separately for Self and Other as well as both Ability conditions, resulting in four additional model parameters. The linear (LD) and normal decay (ND; values depicted in Supplementary Figure 3.2) weighted by the weighting factor *w* that reduced the learning rates towards the ends of the scale were introduced in the learning models in the following way:

[2] $EXP_{t+1} = EXP_t + \alpha\ PE_t\ (1 - w\ LD)$; for the linear decrease;

[3] $EXP_{t+1} = EXP_t + \alpha\ PE_t\ (1 - w\ ND)$; for the normal decrease.

In contrast to our previous studies in which we implemented the LOOP task with fixed feedback sequences, here, feedback depended on the participants' current expectations and thus differed between participants and conditions. Reduced learning rates towards the ends of the feedback scale, which may systematically confound learning rates between participants and conditions, were thus accounted for in Models 7 and 8 (see Supplementary Figure 3.2). To test whether the participants' performance expectation ratings can be better explained in terms of PE learning as compared to stable assumptions in each Ability condition, we included a simple Mean Model, with a mean value for each task condition (Model 9).

Model fitting. For model fitting, we used the RStan package (Stan Development Team, 2016. RStan: the R interface to Stan. R package version 2.14.1.), which implements Markov chain Monte Carlo (MCMC) sampling algorithms. All of the learning models in the model space were fitted for each participant individually, and posterior parameter distributions were sampled for each participant. A total of 2400 samples were drawn after 1000 burn-in samples (overall 3400 samples; thinned with a factor of 3) in three MCMC chains. We assessed whether MCMC chains converged to the target distributions by inspecting $\hat{R}$ values for all model parameters (Gelman & Rubin, 1992). Effective sample sizes (n_{eff}) of model parameters, which are estimates of the effective number of independent draws from the posterior distribution, were typically greater than 1500 (for most parameters and subjects). Posterior distributions for all parameters for each of the participants were summarized by their mean as the central tendency, resulting in a single parameter value per participant that we used in order to calculate group statistics.

Bayesian model selection and family inference. For model selection, we estimated pointwise out-of-sample prediction accuracy for all fitted models separately for each participant by approximating leave-one-out cross-validation (LOO; corresponding to leave-one-trial-out per subject; Acerbi et al., 2018; Vehtari et al., 2016). To do so, we applied Pareto smoothed importance sampling (PSIS) using the log-likelihood calculated from the posterior simulations of the parameter values as implemented by Vehtari et al. (2016). Sum PSIS-LOO scores for each model as well as information about $\hat{k}$ values – the estimated shape parameters of the generalized Pareto distribution – indicating the reliability of the PSIS-LOO estimate are depicted in Supplementary Table 3.1. As summarized in Supplementary Table 3.1, very few trials resulted in insufficient parameter values for $\hat{k}$ and thus potentially unreliable PSIS-LOO scores (on average 1.1 trials per subject with $\hat{k} > 0.7$ for the winning model; Vehtari et al., 2016). BMS on PSIS-LOO scores was performed on the group level, accounting for group heterogeneity in the model that best describes learning behavior (Rigoux et al., 2014). This procedure provides the protected exceedance probability for each model (*pxp*), indicating how likely a given model is to have a higher probability of explaining the data than all other models in the comparison set. The Bayesian omnibus risk (*BOR*) indicates the posterior probability that model frequencies for all models are all equal to each other (Rigoux et al., 2014). We also provide difference scores of PSIS-LOO in contrast to the model that won the BMS, which

can be interpreted as a simple 'fixed-effect' model comparison (see Supplementary Table 3.1; Acerbi et al., 2018; Vehtari et al., 2016). Model comparisons according to PSIS-LOO difference scores were qualitatively comparable to the BMS analyses for our data. Posterior predictive checks were conducted following model selection by quantifying whether the predicted data could capture the variance in performance expectation ratings for each subject within each of the experimental conditions using regression analyses. Additionally, to assess whether the winning model captured the core effects in the behavioral data, we repeated the model-free analysis, which we had conducted on the behavioral data, with the data predicted by the winning model (see Supplementary Note 3.2 for results).

Statistical analyses of learning parameters. Model parameters, i.e. learning rates, of the winning models for all experiments were analyzed on the group level. A repeated measures ANOVA was calculated on the learning rates with the factor Agent (Self [$\alpha_{Self/PE+}$, $\alpha_{Self/PE-}$] vs. Other [$\alpha_{Other/PE+}$, $\alpha_{Other/PE-}$]) and factor Prediction Error Sign (PE+ [$\alpha_{Self/PE+}$, $\alpha_{Other/PE+}$] vs. PE- [$\alpha_{Self/PE-}$, $\alpha_{Other/PE-}$]; in line with the winning model, the term bias corresponds to the categorical distinction between feedback with positive PEs vs negative PEs) as well as Group as a between-subject factor (fMRI vs. behavior), testing whether the formation of self-efficacy beliefs was more valence-specific than forming beliefs about another person's performance.
To associate learning biases with self-conscious affect, that is, embarrassment and pride, as well as self-esteem (SDQ-III subscale scores), we calculated a normalized learning rate valence bias score for self-related learning (Valence Learning Bias = ($\alpha_{Self/PE+}$ - $\alpha_{Self/PE-}$)/ ($\alpha_{Self/PE+}$ + $\alpha_{Self/PE-}$); Müller-Pinzler et al., 2019; Niv et al., 2012; Palminteri et al., 2017). Spearman correlations were calculated between Valence Learning Bias, affect ratings, and self-esteem scores. Statistical tests were performed two-sided if not mentioned otherwise. All statistical analyses on the behavioral data apart from the modeling procedure were performed using *jamovi* (Version 1.2.27, The jamovi project (2020), retrieved from *https://www.jamovi.org*).

Pupil data analysis. For the fMRI sample, eye-tracking data were assessed during scanning. Pupil diameter and gaze behavior were recorded non-invasively in one eye at 500 Hz using an MRI-compatible Eyelink-1000 plus device (SR Research, Kanata, ON, Canada) with manufacturer-recommended settings for calibration and blink detection. Stimuli were presented on a TFT display (32'', Active area 698.4 mm(H) x 392.85 mm(V); Pixels 1920 x 1080; NordicNeuroLab's LCD monitor [NNL MRI InroomViewingDevice; NordicNeuroLab AS, Møllendalsveien 65 C, 5009 Bergen, Norway]), located 50 cm from the observer, in an otherwise dark room. The task has been optimized with respect to eye-tracking by controlling the global luminance of the stimuli as well as the local luminance of the feedback scale. Due to insufficient pupillometry data quality, three participants had to be excluded from the analyses (final sample n = 36). Pupil data were preprocessed by cutting out periods of blinks, and values in this gap were interpolated by piecewise cubic interpolation. The pupil trace was subsequently z-normalized over the whole session. To

characterize the pupil dilation for each trial by a single value, we calculated a linear slope for each feedback phase of three seconds. Summarizing the pupil dynamics during a single trial with a linear slope is a robust and valid measure for an arousal related pupil response to stimuli of comparable lengths (Krach et al., 2015; Müller-Pinzler et al., 2015; Paulus et al., 2015) building up after 1-2 s until reaching a plateau after more than approximately 6 seconds (Bradley et al., 2008; Geuter et al., 2014). Pupil traces were only analyzed for the Self condition as there was an offset in the pupil diameter at the beginning of the feedback presentation (see Supplementary Figure 3.6). The strong difference in the pupil diameter between the Agent conditions is expected given the greater arousal (Joshi & Gold, 2020) after the cognitive effort when estimating properties. While the pupil slope is a robust measure of rather sustained relative change during stimulus presentation, this offset in the diameter at the beginning of the feedback presentation makes it impossible to draw valid comparisons of these slopes between the Agent conditions as greater pupil diameter will result in greater negative slopes compared to smaller pupil diameter. The linear mixed models (LMMs) with pupil slopes as dependent variable were fitted including intercept, PE valence, and PE surprise both as fixed and random effects. In three separate models either embarrassment ratings, pride ratings, or the Valence Learning Bias (Covariates) were included as second-level covariates as well as their interaction with PE valence (see Supplementary Note 3.6, Supplementary Figures 3.7 - 3.9 for supporting analyses on the linear mixed models). The model description of the full model was as following:

[4] $$s_{i,j} = \gamma_{0,0} + \gamma_{1,0} * PE\ Valence_{i,j} + \gamma_{2,0} * PE\ Surprise_{i,j} + \gamma_{0,1} * Covariate_j + \gamma_{1,1} * Covariate_j * PE\ Valence_{i,j} + \upsilon_{0,j} + \upsilon_{1,j} * PE\ Valence_{i,j} + \upsilon_{2,j} * PE\ Surprise_{i,j} + \varepsilon_{i,j}$$

fMRI data acquisition. Participants were scanned using a 3T Siemens MAGENTOM Skyra scanner (Siemens, München, Germany) at the Center of Brain, Behavior, and Metabolism (CBBM) at the University of Lübeck, Germany with 60 near-axial slices. An echo planar imaging (EPI) sequence was used for the acquisition of on average 1520 functional volumes (min = 1395, max = 1672) during each of the two sessions of the experiment, resulting in a total of on average 3040 functional volumes (TR = 0.992s, TE = 28 ms, flip angle = 60°, voxel size = 3×3×3 mm^3, simultaneous multi-slice factor 4). In addition, a high-resolution anatomical T1 image was acquired, which was used for normalization (voxel size = 1×1×1 mm^3, 192×320×320 mm^3 field of view, TR = 2.300s, TE = 2.94ms, TI = 900ms; flip angle = 9°; GRAPPA factor 2; acquisition time 6.55 min; see Supplementary Figure 3.10 for whole brain mask).

FMRI data analyses. FMRI data were analyzed using SPM12 (www.fil.ion.ucl.ac.uk/spm). Field maps were reconstructed to obtain voxel displacement maps (VDMs). EPIs were corrected for timing differences of the slice acquisition, motion-corrected and unwarped using the corresponding VDMs to correct for geometric distortions and normalized using the forward deformation fields as obtained from the unified segmentation of the anatomical T1 image. The normalized volumes were resliced with a voxel size of 2×2×2 mm^3 and

smoothed with an 8 mm full-width-at-half-maximum isotropic Gaussian kernel. To remove low-frequency drifts, functional images were high-pass filtered at 1/384.

Statistical analyses were performed using a two-level, mixed-effects procedure. A main GLM was implemented on the first level and this fixed-effects GLM included four epoch regressors modeling the hemodynamic responses to the different cue conditions (Ability: High vs. Low × Agent: Self vs. Other), weighted with the performance expectation ratings per trial as parametric modulator for each condition. Two regressors modeled the feedback conditions for Self vs Other collapsing across PE valence (Agent: Self vs. Other). Two parametric modulators were included per feedback condition, weighting feedback trials with PE valence (continuous effect of the signed PE values) and PE surprise (continuous effect of the unsigned PE values). Parametric modulators were not orthogonalized, thus each only explaining their specific variance. One regressor modeled the performance expectation rating phase. The estimation periods for Self and Other were modeled as two regressors, and emotion rating phases as separate regressor. Each of the regressors was modeled with the exact duration as presented during the experiment: The cue phase was modeled with a duration of 2.5 s, the expectation rating phase according to individual reaction times with a mean of 4.26 s (SD = 1.04), the estimation phase with 10 s, the feedback phase with 3 s, and the emotion rating phase with 22.51 s (SD = 3.85). To account for noise due to head movement, six additional regressors modeling head movement parameters were introduced and a constant term was included for each of the two sessions.

On the second level, beta images for the parametric weights of feedback were extracted for Self and Other. Four separate one sample t-tests were implemented for PE valence and PE surprise for Self and Other. For direct comparisons of PE valence and PE surprise responses for Self and Other, two repeated measures ANOVAs were conducted including the respective beta images for Self and Other. Differential tracking of the PE valence, depending on biased learning and self-conscious affect, were examined by three additional second-level models for the PE valence beta images for Self, including either the Valence Learning Bias, embarrassment, or pride ratings as between-subject covariate. A self-related Pupil Dilation Bias (average slope for positive PEs - average slope for negative PEs; higher scores indicate stronger pupil dilation for positive PEs) was also included as covariate in another second-level model to assess whether the neural response scaling with more negative PEs was associated with the pupil dilation response. Here, we tested for stronger responses with more negative PEs associated with more negative affect and a more negative Valence Learning Bias and Pupil Dilation Bias. The analyses including all covariates were conducted within our predefined ROIs, the bilateral dAI, vAI, amygdala, mPFC, and VTA/ SN, as these regions are associated with affective and motivational aspects and PE tracking during learning (for a detailed description see 3.5 Methods - Thresholding procedure and regions of interest). Supplementary analyses assessing all four feedback conditions and parametric modulators are show in Supplementary Note 3.4, Supplementary Figure 3.11, Supplementary Figure 3.12, Supplementary Figure 3.13.

We additionally performed psychophysiological interaction (PPI) analyses on the first level, investigating whether functional connectivity of the dAI, which is commonly activated during feedback processing independent of Agent and Prediction Error Sign (conjunction of baseline contrasts: feedback Self ∩ feedback Other), would differ depending on the PE valence. PPI analyses were computed separately for Self and Other and the resulting contrast images for the PPI effects were aggregated on the second level using two-sample t-tests contrasting PPI effects for Self vs. Other. For each participant, we defined 6-mm radius spherical ROIs, centered at the nearest local maximum for the conjunction contrast feedback Self ∩ feedback Other and located within 10 mm of the group maximum within the dAI, separately for the left dAI (x, y, z: -33 20 -4) and right dAI (x, y, z: 36 20 -7). By computing the first eigenvariate for all voxels within these ROIs that showed a positive effect for the conjunction ($p < .500$), we extracted the time course of activations and constructed PPI terms using the contrast for the parametric weights of PE valence for Self or Other, respectively, resulting in four distinct PPI first-level GLMs. One participant was excluded from the PPI analyses for the right dAI, because no voxels survived the predefined threshold for eigenvariate extraction. The PPI term, along with the activation time course from the (left or right) dAI was included in a new GLM for each participant that also included all the regressors in the initial first-level GLM (four regressors for the different cue conditions, each weighted with the expected performance ratings; two feedback regressors for Self and Other with each two parametric modulators for PE valence and PE surprise; two regressors for the estimation periods for Self and Other; one regressor for the expectation ratings phase; one regressor for the emotion ratings phase; six regressors modeling head movement parameters; a constant term for each session). On the second level, we assessed whether there was a stronger functional coupling of the dAI seed regions with the predefined ROIs (amygdala, mPFC, VTA/ SN) for the Self in contrast to the Other when PE valence was more negative. In line with the negative Valence Learning Bias for self-efficacy beliefs and stronger pupil dilation responses scaling with more negative PEs we specifically tested for stronger functional connectivity correlated with more negative PEs. Functional connectivity dynamics were also associated with learning behavior by calculating Spearman correlations for the Valence Learning Bias and the mean parameter estimates for the PPI effect of Self > Other extracted from the GLMs described above in a sphere of 6 mm around the peak voxels within the predefined ROIs (amygdala, mPFC, VTA/ SN).

Thresholding procedure and regions of interest. According to its suggested role as an integrative hub for motivated cognition and emotional behavior, the AI was defined as one of the regions of interest (ROIs; Koban & Pourtois, 2014; Wager & Barrett, 2017). Due to their specific functional associations, a bilateral ventral and a bilateral dorsal AI ROI was defined according to the three-cluster solution of Kelly and colleagues (2012). The bilateral amygdala was defined as another ROI and derived from the AAL atlas definition in the WFU PickAtlas (Tzourio-Mazoyer et al., 2002) due to its similar role for the attention-emotion interaction (Kaspar & König, 2012; Koban & Pourtois, 2014). The mPFC ROI was also derived from the AAL atlas in the WFU PickAtlas (label: bilateral frontal superior

medial) due to its specific role during social learning and for biases in self-related learning reported in previous studies (Kuzmanovic et al., 2018; Sharot, 2011). Additionally, an anatomically defined VTA/ SN ROI, dopaminergic nuclei in the midbrain, was included (probabilistic atlases of the midbrain; Adcock Lab; Ballard et al., 2011; Murty et al., 2014) as dopamine signals motivationally important events, e.g. during reward learning (Schultz, 1998), and has been associated with biases in memory towards events that are of motivational relevance (Adcock et al., 2006).

FMRI results were family-wise-error (FWE) corrected at peak level for the whole brain unless ROI analyses were conducted, and all coordinates are reported in MNI space. As our predefined ROIs were chosen with respect to their involvement with the emotion-cognition link, we tested the effects of our covariates on PE valence tracking and PPI effects within the ROIs. Anatomical labels of all resulting clusters were derived from the Automated Labeling Atlas Version 3.0 (Eickhoff et al., 2005).

3.6 References

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS Computational Biology, 14*(7), e1006110. https://doi.org/10.1371/journal.pcbi.1006110

Adcock, R. A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. E. (2006). Reward-motivated learning: mesolimbic activation precedes memory formation. *Neuron*, *50*(3), 507–517. https://doi.org/10.1016/j.neuron.2006.03.036

Alden, L. E., Taylor, C. T., Mellings, T. M. J. B., & Laposa, J. M. (2008). Social anxiety and the interpretation of positive social events. *Journal of Anxiety Disorders*, *22*(4), 577–590. https://doi.org/10.1016/j.janxdis.2007.05.007

Amir, N., Prouvost, C., & Kuckertz, J. M. (2012). Lack of a Benign Interpretation Bias in Social Anxiety Disorder. *Cognitive Behaviour Therapy*, *41*(2), 119–129. https://doi.org/10.1080/16506073.2012.662655

Apsler, R. (1975). Effects of embarrassment on behavior toward others. *Journal of Personality and Social Psychology*, *32*(1), 145–153. https://doi.org/10.1037/h0076699

Ballard, I. C., Murty, V. P., Carter, R. M., Macinnes, J. J., Huettel, S. A., & Adcock, R. A. (2011). Dorsolateral Prefrontal Cortex Drives Mesolimbic Dopaminergic Regions to Initiate Motivated Behavior. *The Journal of Neuroscience*, *31*(28), 10340–10346. https://doi.org/10.1523/JNEUROSCI.0895-11.2011

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Bromberg-Martin, E. S., & Sharot, T. (2020). The Value of Beliefs. *Neuron*, *106*(4), 561–565. https://doi.org/10.1016/j.neuron.2020.05.001

Cecchi, R., Vinckier, F., Hammer, J., Marusic, P., Nica, A., Rheims, S., Trebuchon, A., Barbeau, E. J., Denuelle, M., Maillard, L., Minotti, L., Kahane, P., Pessiglione, M., & Bastin, J. (2022). Intracerebral mechanisms explaining the impact of incidental feedback on mood state and risky choice. *ELife*, *11*. https://doi.org/10.7554/eLife.72440

Charpentier, C. J., Bromberg-Martin, E. S., & Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(31), E7255–E7264. https://doi.org/10.1073/pnas.1800547115

Charpentier, C. J., De Martino, B., Sim, A. L., Sharot, T., & Roiser, J. P. (2016). Emotion-induced loss aversion and striatal-amygdala coupling in low-anxious individuals. *Social Cognitive and Affective Neuroscience*, *11*(4), 569–579. https://doi.org/10.1093/scan/nsv139

Charpentier, C. J., De Neve, J. E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of Affective Decision Making: How Do Feelings Predict Choice? *Psychological Science*, *27*(6), 763–775. https://doi.org/10.1177/0956797616634654

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, *29*(39), 12315–12320. https://doi.org/10.1523/JNEUROSCI.2575-09.2009

Christianson, S. A. (2014). *The Handbook of Emotion and Memory: Research and Theory* (S. A. Christianson, Ed.). Taylor & Francis.

Craig, A. D. (2003). Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*(4), 500–505. https://doi.org/10.1016/s0959-4388(03)00090-4

Czekalla, N., Stierand, J., Stolz, D. S., Mayer, A. V., Voges, J. F., Rademacher, L., Paulus, F. M., Krach, S., & Müller-Pinzler, L. (2021). Self-beneficial belief updating as a coping mechanism for stress-induced negative affect. *Scientific Reports*, *11*(1), 17096. https://doi.org/10.1038/s41598-021-96264-0

Darby, R. S., & Harris, C. R. (2010). Embarrassment's effect on facial processing. *Cognition and Emotion*, *24*(7), 1250–1258. https://doi.org/10.1080/02699930903211183

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences*, *111*(5), E618–E625. https://doi.org/10.1073/pnas.1317557111

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. *Social Cognitive and Affective Neuroscience*, *12*(4), 618–634. https://doi.org/10.1093/scan/nsw171

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034

Elder, J., Davis, T., & Hughes, B. L. (2022). Learning About the Self: Motives for Coherence and Positivity Constrain Learning From Self-Relevant Social Feedback. *Psychological Science*, *33*(4), 629–647. https://doi.org/10.1177/09567976211045934

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545. https://doi.org/10.1016/j.jebo.2011.05.012

Feinberg, M., Willer, R., & Keltner, D. (2012). Flustered and faithful: embarrassment as a signal of prosociality. *Journal of Personality and Social Psychology*, *102*(1), 81–97. https://doi.org/10.1037/a0025403

Frijda, N. H. (1987). Emotion, cognitive structure, and action tendency. *Cognition and Emotion*, *1*(2), 115–143. https://doi.org/10.1080/02699938708408043

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating Beliefs under Perceived Threat. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.1523/JNEUROSCI.0716-18.2018

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Geuter, S., Gamer, M., Onat, S., & Büchel, C. (2014). Parametric trial-by-trial prediction of pain by easily available physiological measures. *Pain*, *155*(5), 994–1001. https://doi.org/10.1016/j.pain.2014.02.005

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, *28*(22), 5623–5630. https://doi.org/10.1523/JNEUROSCI.1309-08.2008

Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2010). A cognitive- behavioral model of social anxiety disorder: Update and extension. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social anxiety: Clinical, developmental, and social perspectives* (pp. 395–422). NY: Elsevier.

Hopkins, A. K., Dolan, R., Button, K. S., & Moutoussis, M. (2021). A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious Individuals. *Computational Psychiatry*, *5*(1), 21. https://doi.org/10.5334/cpsy.57

Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19*(2), 62–64. https://doi.org/10.1016/j.tics.2014.12.006

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of Cognitive Neuroscience*, *22*(4), 621–631. https://doi.org/10.1162/jocn.2009.21228

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. *Trends in Cognitive Sciences*, *24*(6), 466–480. https://doi.org/10.1016/j.tics.2020.03.005

Kaspar, K., & König, P. (2012). Emotions and personality traits as high-level factors in visual attention: a review. *Frontiers in Human Neuroscience*, *6*, 1–14. https://doi.org/10.3389/fnhum.2012.00321

Kelly, C., Toro, R., Di Martino, A., Cox, C. L., Bellec, P., Castellanos, F. X., & Milham, M. P. (2012). A convergent functional architecture of the insula emerges across imaging modalities. *NeuroImage*, *61*(4), 1129–1142. https://doi.org/10.1016/j.neuroimage.2012.03.021

Keltner, D., & Potegal, M. (1997). Appeasement and reconciliation: Introduction to an aggressive behavior special issue. *Aggressive Behavior*, *23*(5), 309–314. https://doi.org/10.1002/(SICI)1098-2337(1997)23:5<309::AID-AB1>3.0.CO;2-D

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neuroscience and Biobehavioral Reviews*, *46 Pt 1*(1), 71–84. https://doi.org/10.1016/j.neubiorev.2014.02.014

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., Wager, T. D., & Arch, J. J. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion* , *17*(8), 1144–1155. https://doi.org/10.1037/emo0000296

Koenig, S., Uengoer, M., & Lachnit, H. (2018). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology*, *55*(4), 1–12. https://doi.org/10.1111/psyp.13020

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592. https://doi.org/10.1017/S0033291713001074

Korn, Christoph W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Krach, S., Kamp-Becker, I., Einhäuser, W., Sommer, J., Frässle, S., Jansen, A., Rademacher, L., Müller-Pinzler, L., Gazzola, V., & Paulus, F. M. (2015). Evidence from pupillometry and fMRI indicates reduced neural response during vicarious social pain but not physical pain in autism. *Human Brain Mapping*, *36*(11), 4730–4744. https://doi.org/10.1002/hbm.22949

Kube, T., Kirchner, L., Lemmer, G., & Glombiewski, J. A. (2021). How the Discrepancy Between Prior Expectations and New Information Influences Expectation Updating in Depression—The Greater, the Better? *Clinical Psychological Science*. https://doi.org/10.1177/21677026211024644

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kurth, F., Zilles, K., Fox, P. T., Laird, A. R., & Eickhoff, S. B. (2010). A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure & Function*, *214*(5–6), 519–534. https://doi.org/10.1007/s00429-010-0255-z

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162. https://doi.org/10.1016/j.neuroimage.2016.02.014

Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(37), 7996–8010. https://doi.org/10.1523/JNEUROSCI.0266-18.2018

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences*, *35*(3), 121–143. https://doi.org/10.1017/S0140525X11000446

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(35), 9763–9768. https://doi.org/10.1073/pnas.1603198113

Lockwood, P. L., & Wittmann, M. K. (2018). Ventral anterior cingulate cortex and social decision-making. *Neuroscience and Biobehavioral Reviews*, *92*, 187–191. https://doi.org/10.1016/j.neubiorev.2018.05.030

Loewenstein, G. (2006). Social science. The pleasures and pains of information [Review of *Social science. The pleasures and pains of information*]. *Science*, *312*(5774), 704–706. https://doi.org/10.1126/science.1128388

Loued-Khenissi, L., Pfeuffer, A., Einhäuser, W., & Preuschoff, K. (2020). Anterior insula reflects surprise in value-based decision-making and perception. *NeuroImage*, *210*, 116549. https://doi.org/10.1016/j.neuroimage.2020.116549

Markus, H., & Wurf, E. (1987). The Dynamic Self-Concept: A Social Psychological Perspective. *Annual Review of Psychology*, *38*(1), 299–337. https://doi.org/10.1146/annurev.psych.38.1.299

Marsh, H. W., & O'Neill, R. (1984). Self Description Questionnaire III: The Construct Validity of Multidimensional Self-Concept Ratings by Late Adolescents. *Journal of Educational Measurement*, *21*(2), 153–174.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure & Function*, *214*(5–6), 655–667. https://doi.org/10.1007/s00429-010-0262-0

Mesulam, M. M., & Mufson, E. J. (1982). Insula of the old world monkey. I. Architectonics in the insulo-orbito-temporal component of the paralimbic brain. *The Journal of Comparative Neurology*, *212*(1), 1–22. https://doi.org/10.1002/cne.902120102

Miller, R. S. (1996). *Embarrassment: Poise and peril in everyday life*. Guilford Press.

Morrison, A. S., & Heimberg, R. G. (2013). Social anxiety and social anxiety disorder. *Annual Review of Clinical Psychology*, *9*, 249–274. https://doi.org/10.1146/annurev-clinpsy-050212-185631

Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., Einhäuser, W., Paulus, F. M., & Krach, S. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, *119*, 252–261. https://doi.org/10.1016/j.neuroimage.2015.06.036

Müller-Pinzler, Laura, Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, *9*(1), 14416. https://doi.org/10.1038/s41598-019-50821-w

Murty, V. P., Shermohammed, M., Smith, D. V., Carter, R. M., Huettel, S. A., & Adcock, R. A. (2014). Resting state networks distinguish human ventral tegmental area from substantia nigra. *NeuroImage*, *100*, 580–589. https://doi.org/10.1016/j.neuroimage.2014.06.047

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/jneurosci.5498-10.2012

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, *14*(6), 769–776. https://doi.org/10.1016/j.conb.2004.10.016

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Paulus, F. M., Krach, S., Blanke, M., Roth, C., Belke, M., Sommer, J., Müller-Pinzler, L., Menzler, K., Jansen, A., Rosenow, F., Bremmer, F., Einhäuser, W., & Knake, S. (2015). Fronto-insula network activity explains emotional dysfunctions in juvenile myoclonic epilepsy: combined evidence from pupillometry and fMRI. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *65*, 219–231. https://doi.org/10.1016/j.cortex.2015.01.018

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*, *16*(2), 331–348. https://doi.org/10.1006/nimg.2002.1087

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Frontiers in Neuroscience*, *5*(September), 115. https://doi.org/10.3389/fnins.2011.00115

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-09075-3

Rouhani, N., & Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *ELife*, *10*, e61077. https://doi.org/10.7554/eLife.61077

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews. Neuroscience*, *15*(8), 549–562. https://doi.org/10.1038/nrn3776

Rutledge, R. B., de Berker, A. O., Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*(May), 11825. https://doi.org/10.1038/ncomms11825

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(40), 13525–13536. https://doi.org/10.1523/JNEUROSCI.1747-10.2010

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, *40*. https://doi.org/10.1152/jn.1998.80.1.1

Schultz, W., Dayan, P., & Montague, P. R. (1997). *A Neural Substrate of Prediction and Reward on JSTOR*. *275*(MARCH).

Sedikides, C., & Gregg, A. P. (2008). Self-Enhancement: Food for Thought. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *3*(2), 102–116. https://doi.org/10.1111/j.1745-6916.2008.00068.x

Sedikides, C., & Hepper, E. G. D. (2009). Self-Improvement. *Social and Personality Psychology Compass*, *3*, 899–917. https://doi.org/10.1111/j.1751-9004.2009.00231.x

Sharot, T. (2011). The optimism bias. In *Current Biology* (Vol. 21, Issue 23, pp. R941–R945). Cell Press. https://doi.org/10.1016/j.cub.2011.10.030

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/j.tics.2015.11.002

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. https://doi.org/10.1038/nn.2949

Stolz, D. S., Müller-Pinzler, L., Krach, S., & Paulus, F. M. (2020). Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nature Communications*, *11*(1), 1230. https://doi.org/10.1038/s41467-020-14800-4

Storbeck, J., & Clore, G. L. (2008). Affective Arousal as Information: How Affective Arousal Influences Judgments, Learning, and Memory. *Social and Personality Psychology Compass*, *2*(5), 1824–1843. https://doi.org/10.1111/j.1751-9004.2008.00138.x

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210. https://www.ncbi.nlm.nih.gov/pubmed/3283814

Touroutoglou, A., Hollenbeck, M., Dickerson, B. C., & Feldman Barrett, L. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *NeuroImage*, *60*(4), 1947–1958. https://doi.org/10.1016/j.neuroimage.2012.02.012

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Ullsperger, M., Harsay, H. A., Wessel, J. R., & Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure & Function*, *214*(5–6), 629–643. https://doi.org/10.1007/s00429-010-0261-1

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, *17*(1), 3581–3618.

Vinckier, F., Rigoux, L., Kurniawan, I. T., Hu, C., Bourgeois-Gironde, S., Daunizeau, J., & Pessiglione, M. (2019). Sour grapes and sweet victories: How actions shape preferences. *PLoS Computational Biology*, *15*(1), e1006499. https://doi.org/10.1371/journal.pcbi.1006499

Vinckier, F., Rigoux, L., Oudiette, D., & Pessiglione, M. (2018). Neuro-computational account of how mood fluctuations arise and affect decision making. *Nature Communications*, *9*(1), 1708. https://doi.org/10.1038/s41467-018-03774-z

Wager, T. D., & Barrett, L. F. (2017). From affect to control: Functional specialization of the insula in motivation and regulation. In *bioRxiv* (p. 102368). https://doi.org/10.1101/102368

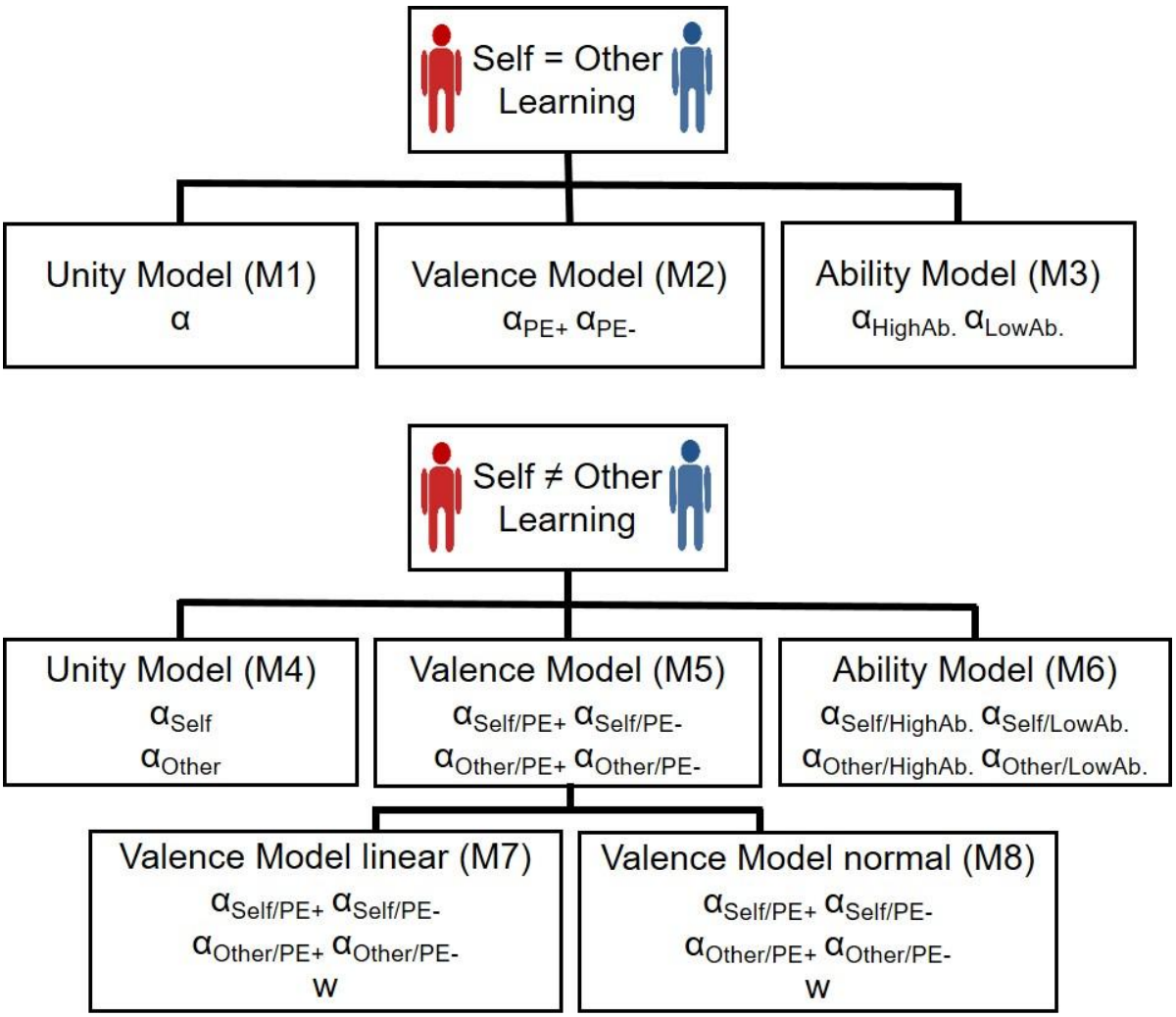
Wallis, J. D., & Kennerley, S. W. (2010). Heterogeneous reward signals in prefrontal cortex. *Current Opinion in Neurobiology*, *20*(2), 191–198. https://doi.org/10.1016/j.conb.2010.02.009

Will, G.-J., Moutoussis, M., Womack, P. M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., NSPN Consortium, Rutledge, R. B., & Dolan, R. J. (2020). Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Translational Psychiatry*, *10*(1), 96. https://doi.org/10.1038/s41398-020-0702-4

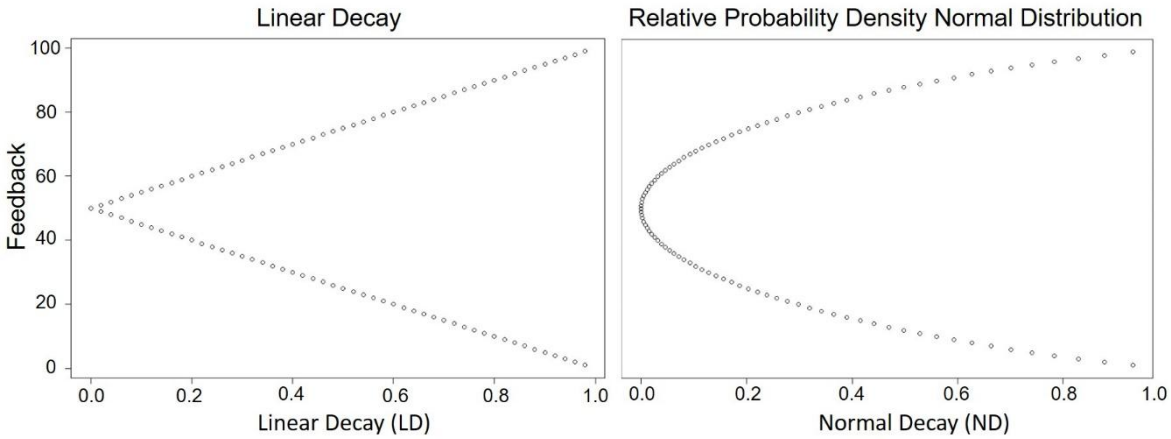
Williams, L. A., & DeSteno, D. (2008). Pride and perseverance: the motivational role of pride. *Journal of Personality and Social Psychology*, *94*(6), 1007–1017. https://doi.org/10.1037/0022-3514.94.6.1007

3.7 Supplementary Information

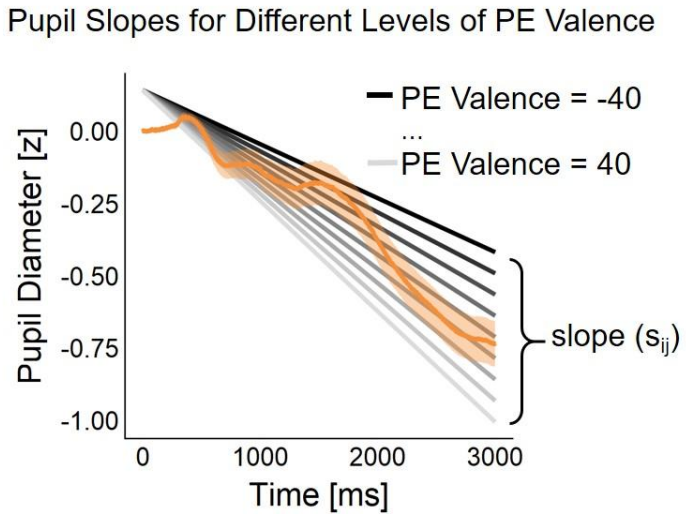
Supplementary Figures



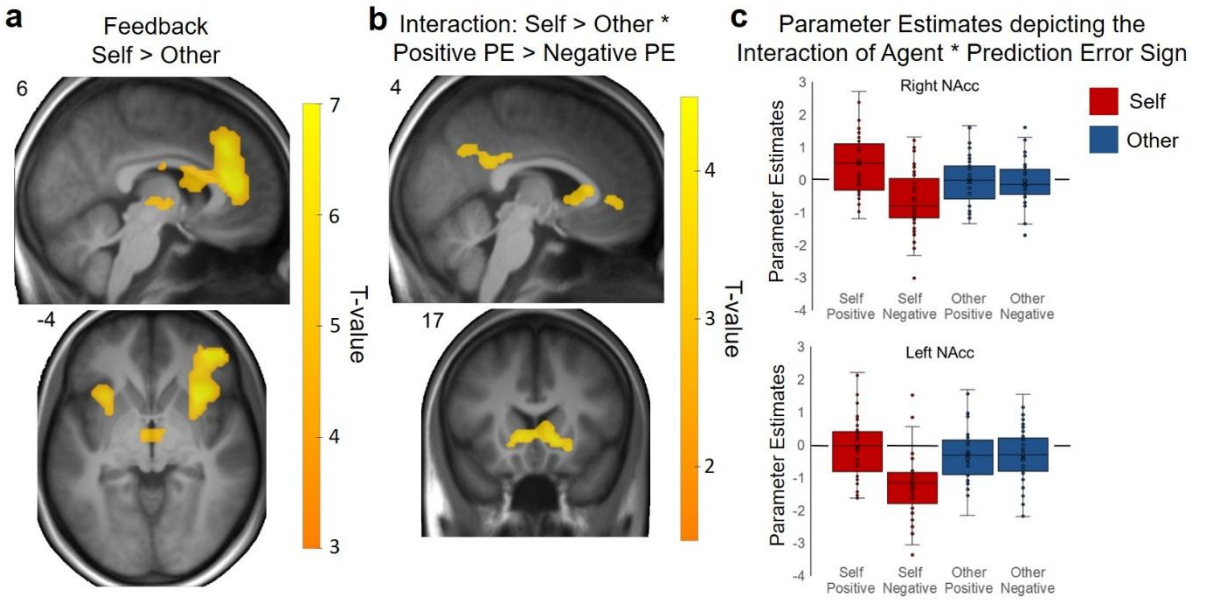
Supplementary Figure 3.1. Structure of the model space. Two factors were distinguished that impact learning rates (α): the agent (self vs other) and the impact (no impact: Unity Model) of prediction error valence (Valence Model) or the ability condition (Ability Model). The Valence Model, winning model in previous studies (Müller-Pinzler et al., 2019), was extended by a decay factor (w) for the learning rates towards the ends of the feedback scale with a linear decrease (Valence Model linear) or a decrease following the relative probability density of the normal distribution (Valence Model normal; for more details see 3.5 Methods).



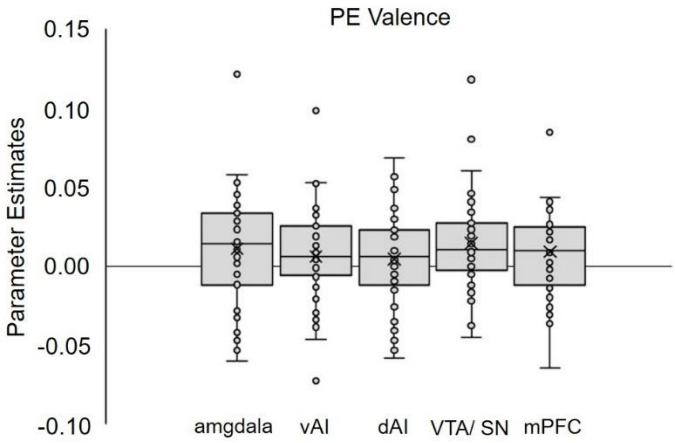
Supplementary Figure 3.2. Depiction of the linear decay (left) and the decay following the relative probability density of the normal distribution (right) for the different feedback values. The values depicted here were introduced in the learning models and weighted by a weighting factor as described in the 3.5 Methods section.



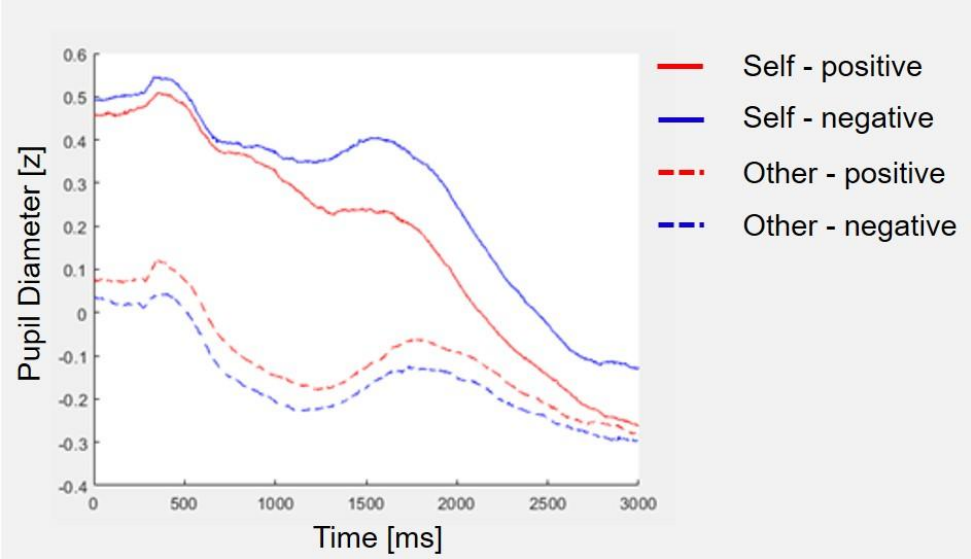
Supplementary Figure 3.3. Average pupil trace and pupil slopes for different levels of PE Valence. More negative PEs are associated with greater pupil slopes. The average pupil diameter trace during feedback is depicted in orange, the shaded area represents +/- one standard error. Pupil slopes for the different levels of PEs (from black = negative to grey = positive) were predicted by the multi-level model containing PE valence and PE surprise as predictors, as described in the 3.5 Methods section.



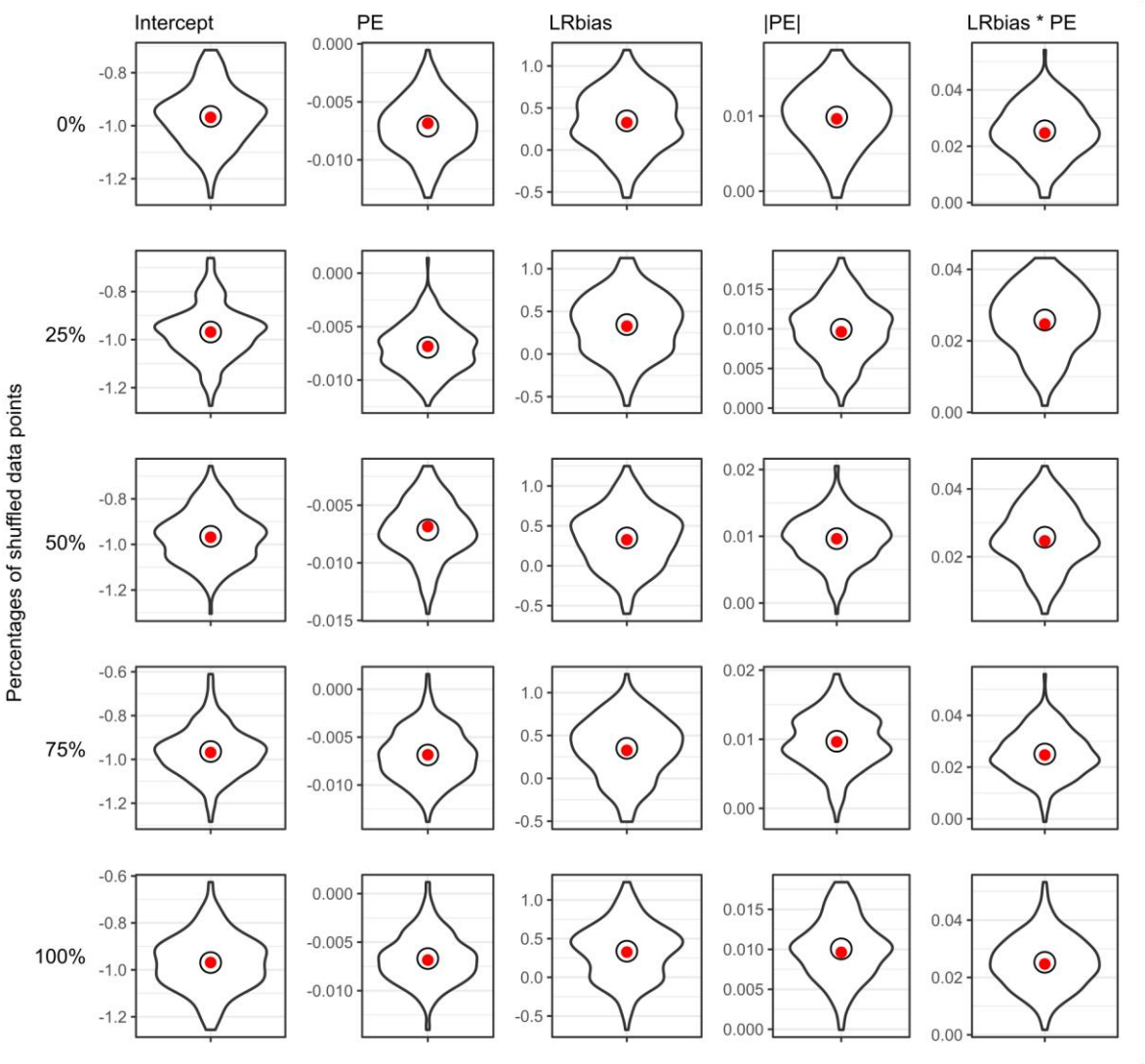
Supplementary Figure 3.4. Neural activations associated with feedback processing. a) Self-related feedback vs. other-related feedback was associated with an increased activation of the mPFC/ ACC, bilateral anterior insula and thalamus, among other regions ($p < .05$, FWE corrected at peak level for the whole brain). **b)** The interaction of Agent and Prediction Error Sign (([Self positive PE > Self negative PE] > [Other positive PE > Other negative PE]) resulted in activation of the angular gyrus, the bilateral NAcc/VS, the precuneus/ posterior cingulate cortex, and precentral gyrus (cluster-wise FWE corrected with $p < .05$ at a cluster forming threshold of $p < .001$ for displaying purposes). **c)** Parameter estimates correspond to the BOLD response to positive and negative PEs in bilateral NAcc/ VS [left: x, y, z: -9 20 -1; right: x, y, z: 12 20 -1]. The plot shows that positive relative to negative PEs increased the activity in NAcc/ VS only when learning about the own performance, but not when observing others. Colored bars indicate the first and third quartile of the data, the line marks the median, the cross marks the mean. Whiskers extend from the upper (lower) box borders to the largest (smallest) data point at most 1.5 times the interquartile range above (below) the respective border. Data with more extreme values than this are displayed as individual points.



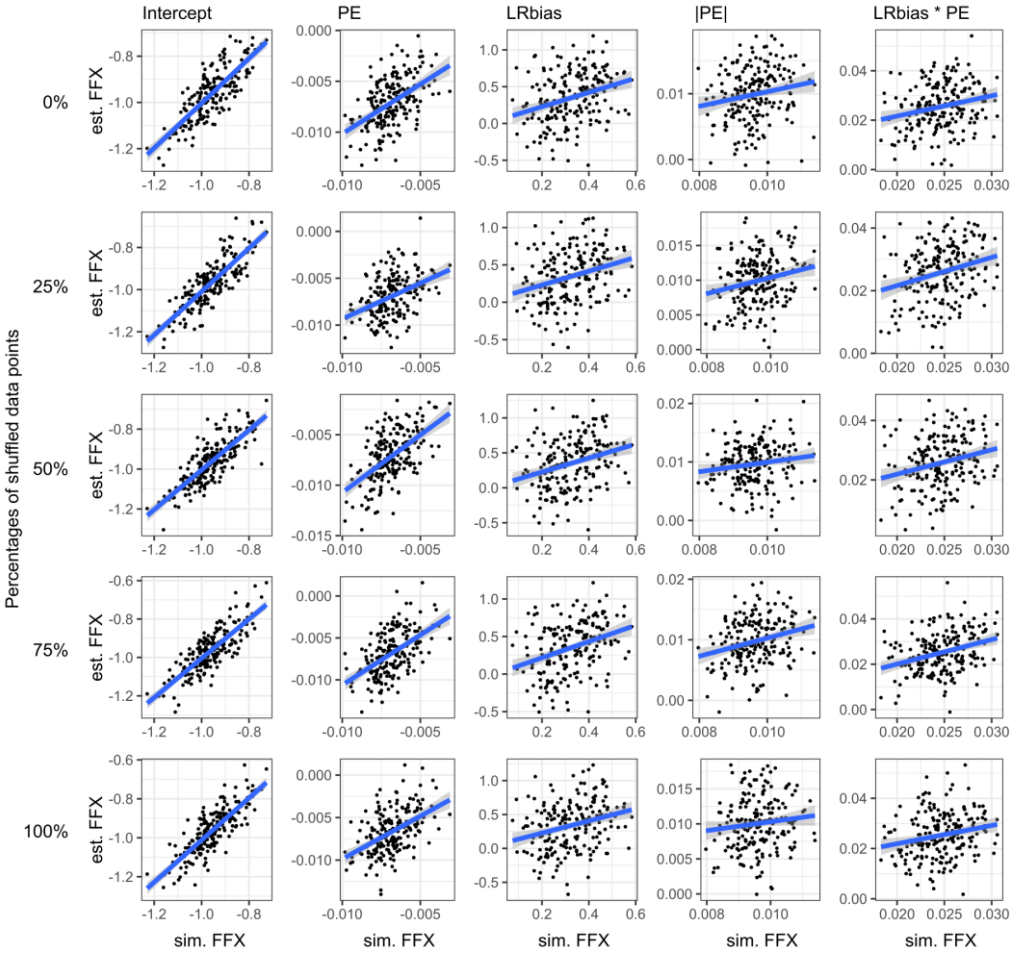
Supplementary Figure 3.5. Parameter estimates for the PE Valence effect in our predefined ROIs (amygdala, vAI, dAI, VTA/SN, and mPFC). Parameter estimates were derived from all voxels within each ROI and averaged across all voxels. Left and right amygdala, vAI, and dAI were combined bilaterally for displaying purposes. Dots show the data for individual subjects, grey bars show the first and third quartile, the cross marks the mean and the line the median. Whiskers extend from the upper (lower) box borders to the largest (smallest) data point at most 1.5 times the interquartile range above (below) the respective border. Data with more extreme values than this are displayed as individual points. All regions show variance between subjects in such a way that some individuals have positive values and other negative values indicating stronger activity scaling with more positive or more negative PEs, respectively.



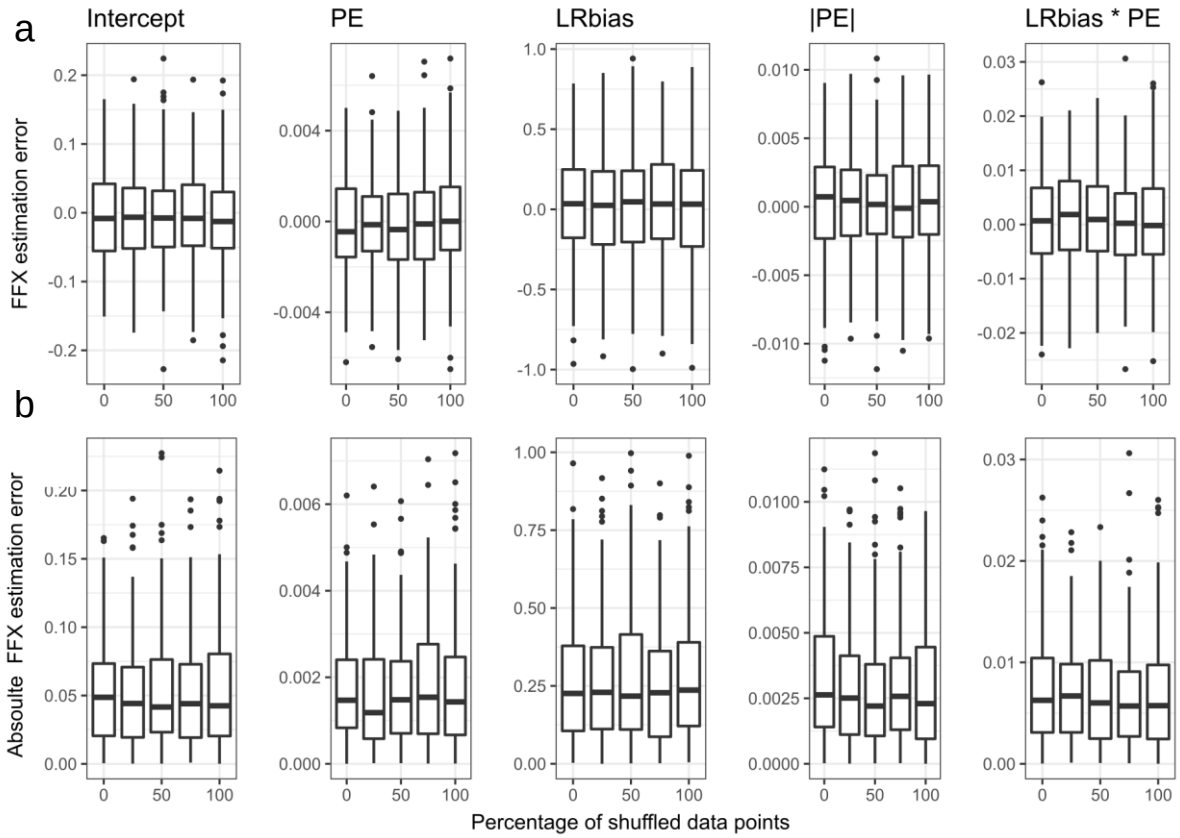
Supplementary Figure 3.6. Pupil diameter traces for the feedback phase (3 secs) for all four feedback conditions uncorrected for offsets at feedback start. The bold lines show pupil traces for Self and the dashed lines for Other. Red lines show pupil traces for feedback with positive PEs and blue lines for negative PEs.



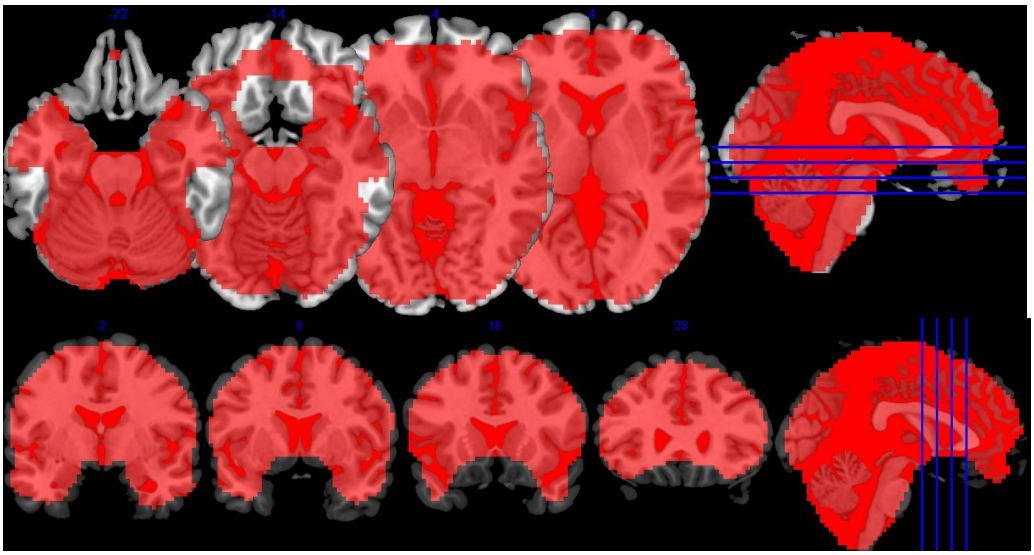
Supplementary Figure 3.7. Visual representation of parameter recovery for mixed model effects in the pupil data. See **Supplementary Note 3.6** for a detailed description. Empirical FFX estimates (as red dots) overlayed on top of violin plots of the recovered estimates; the means of the recovered parameters are displayed as empty black circles



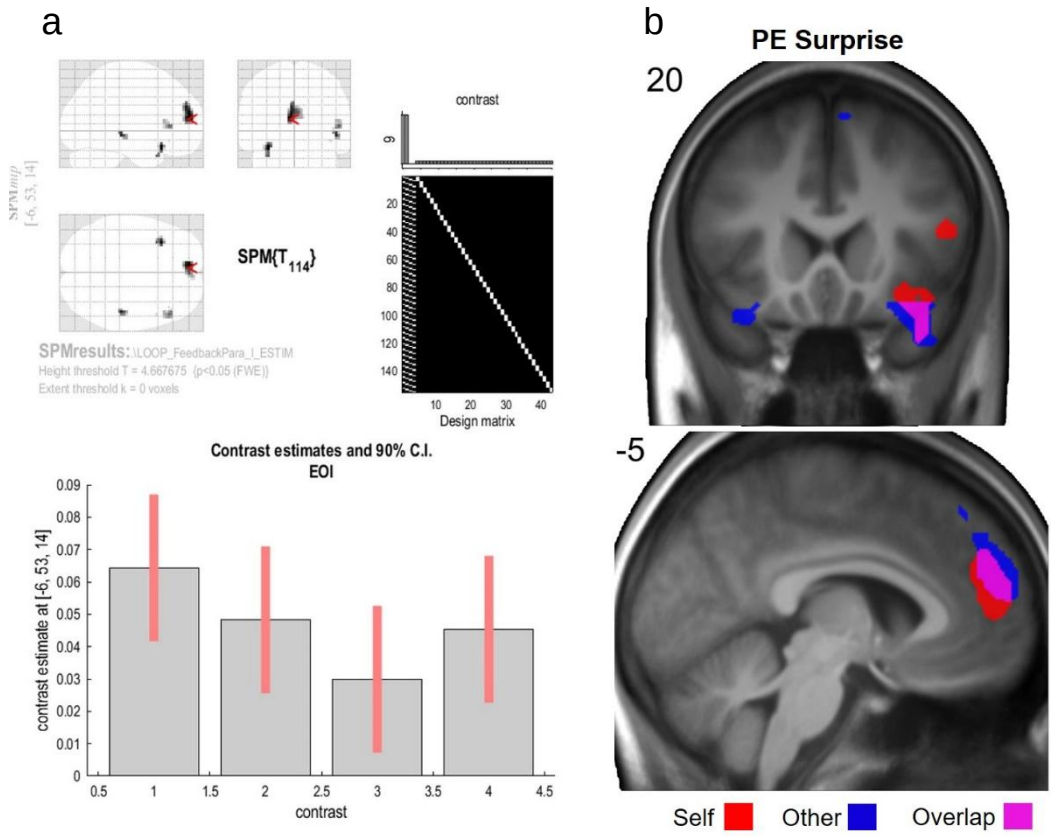
Supplementary Figure 3.8. Correlation plots for parameter recovery analyses. Correlation plots between each set of recovered FFX and the corresponding parameter underlying the simulated data. Blue trend lines are slopes fitted by linear regression and shadings show the 95% confidence intervals for each trend line. See Supplementary Note 3.6 for a detailed description.



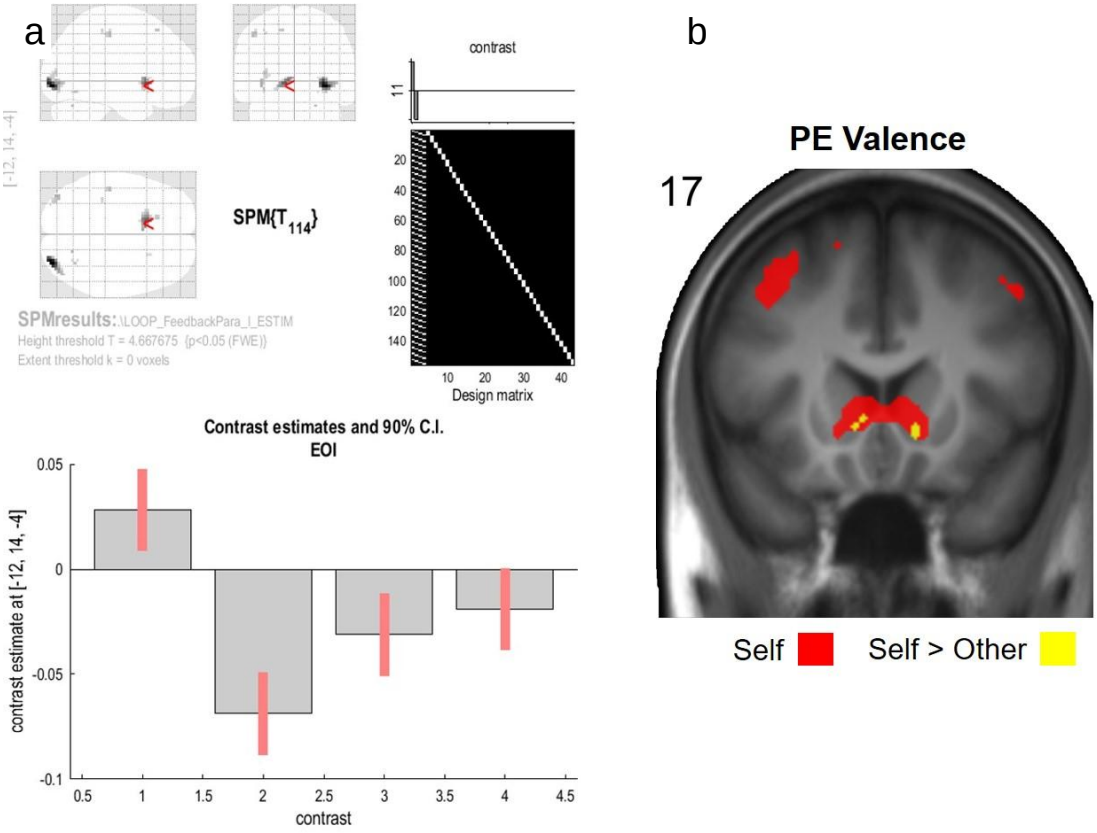
Supplementary Figure 3.9. Differences between recovered and underlying parameters for mixed model effects in the pupil data. a) Box plots for the differences between recovered and underlying parameters (FFX estimation error) for each fixed effect (intercept, signed PE, LRbias, unsigned PE, LRbias * signed PE) and all levels of tiling. **b)** Box plots for the absolute differences between recovered and underlying parameters (absolute FFX estimation error) for each fixed effect (intercept, signed PE, LRbias, unsigned PE, LRbias * signed PE) and all levels of tiling. Lower and upper box borders define the first and third quartile and the thick horizontal line within each box marks the median. Whiskers extend from the upper (lower) box borders to the largest (smallest) data point at most 1.5 times the interquartile range above (below) the respective border. Data with more extreme values than this are displayed as individual points. See Supplementary Note 3.6 for a detailed description.



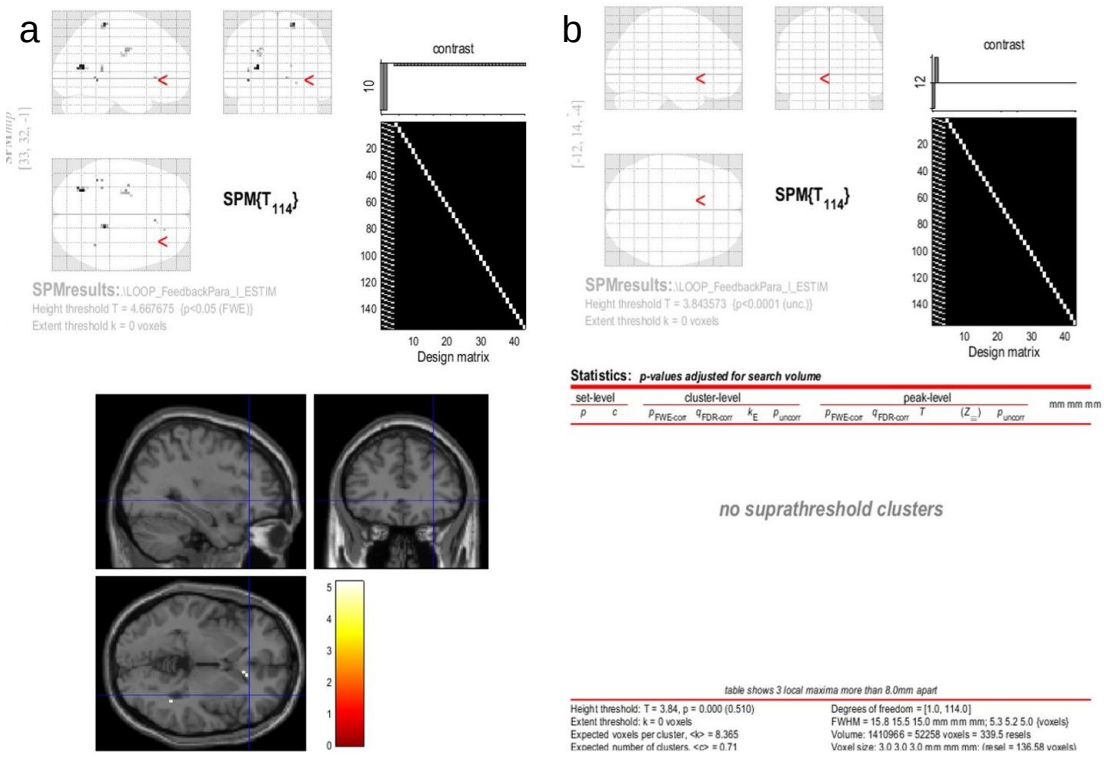
Supplementary Figure 3.10. Brain mask of the 2nd level fMRI analyses. Brain coverage and signal loss was as expected for fMRI studies with preserved signal in our regions of interest. Red shading shows the whole brain mask as derived from the normalized EPI images across subjects, i.e. those voxels that show sufficient signal in all included individuals. Blue lines show the origin of each of the slices depicted within the respective MNI coordinate axis (z above and y below).



Supplementary Figure 3.11. Visual comparison of effects when assessing parametric weights separately or combined for positive vs negative feedback conditions: PE valance effects. a) PE valence effect for Self-Positive > Self-Negative. The flexible factorial model includes parametric weights for PEs for Self-Positive (1), Self-Negative (2), Other-Positive (3), and Other-Negative (4). The contrast PE valence for Self-Positive > Self-Negative is depicted and parameter estimates are shown for the peak voxel in the mPFC. The contrast combines two effects: The effect of PE valence for the Self-Positive contrast entails the PE valence and also the PE surprise effect. The negative PE Valence effect of Self-Negative also entails the positive PE surprise effect. As PE surprise effects add up for the two conditions (Self-Positive and Self-Negative) and PE valence effects are subtracted here, the effect shown is therefore comparable to the PE surprise effect in our original analysis as indicated by the parameter estimates for the mPFC. Results are depicted for illustration purposes and visual comparison with our original analytical approach described in the 3.5 Methods section. The Other conditions are of no interest here. Parametric weights for PEs are coded as unsigned values in the GLM which explains the contrast weights in the Self-Negative condition (i.e. -1 codes more positive PEs). **b)** Shows the corresponding PE surprise effect in our current analyses for visual comparison (see also Figure 3.4).



Supplementary Figure 3.12. Visual comparison of effects when assessing parametric weights separately or combined for positive vs negative feedback conditions: PE surprise effects. a) PE surprise for Self-Positive > Self-Negative. The flexible factorial model includes parametric weights for PEs for Self-Positive (1), Self-Negative (2), Other-Positive (3), and Other-Negative (4). The contrast PE surprise for Self-Positive > Self-Negative is depicted and parameter estimates are shown for the peak voxel in the ventral striatum. The effect of PE surprise Self-Positive again entails PE valence besides the PE surprise effect and the negative PE surprise effect of Self-Negative also entails the positive effect of PE Valence. Here, PE valence effects add up and PE surprise are subtracted. Results are therefore comparable to the PE valence effect in our original analysis as indicated by the parameter estimates for the ventral striatum. Results are depicted for illustration purposes and visual comparison with our original analytical approach described in the 3.5 Methods section. The Other conditions are of no interest here. Parametric weights for PEs are coded as unsigned values which explains the contrast weights in the Self-Negative condition (i.e. -1 codes more positive PEs). **b)** Shows the corresponding PE valence effect in our current analyses (derived from Figure 3.4).



Supplementary Figure 3.13. Effects when assessing parametric weights separately for positive vs negative feedback conditions: Self-Negative > Self-Positive. a) PE valence for Self-Negative > Self-Positive, the opposite comparisons as shown in Supplementary Figure 3.11, did not yield any further effects b) PE surprise for Self-Negative > Self-Positive, the opposite comparisons as shown in Supplementary Figure 3.12, did not yield any further effects. The flexible factorial model includes parametric weights for PEs for Self-Positive (1), Self-Negative (2), Other-Positive (3), and Other-Negative (4). Results are depicted for illustration purposes and visual comparison with our original analytical approach described in the 3.5 Methods section. The Other conditions are of no interest here. Parametric weights for PEs are coded as unsigned values which explains the contrast weights in the Self-Negative condition (i.e. -1 codes more positive PEs).

Supplementary Notes

Supplementary Note 3.1: *Model free behavioral analyses reveal more negative self-evaluation.*

First, we performed a model-free analysis to capture the basic effects observed in our behavioral data. Analyses of behavioral data and learning rates are based on the combined fMRI (n=39) and behavioral sample (n=30; total sample N=69). The Trial x Ability condition x Agent condition linear mixed model revealed a significant main effect of Ability condition (β=-20.54, $t_{(68)}$=-13.45, p<.001, *95% CI*=[-23.54; -17.55]) and interaction of Trial x Ability condition (β=-1.23, $t_{(5240)}$=-36.55, p<.001, *95% CI*=[-1.30; -1.16]), indicating that participants adapted their performance expectation ratings over time according to the feedback provided in each Ability condition (see Figure 3.1c). Moreover, there was a significant main effect of Agent condition (β=6.78, $t_{(68)}$=6.52, p<.001, *95% CI*=[4.74; 8.82]), indicating that participants evaluated their own performance more

negatively than the other's performance. There was also a significant interaction of Agent condition x Ability condition (β=1.29, $t_{(5240)}$=3.33, p<.001, *95% CI*=[0.53; 2.05]). The three-way interaction of Trial x Agent condition x Ability condition (β=0.18, $t_{(5240)}$=2.66, p<.001, *95% CI*=[0.05; 0.31]) revealed a significant effect, hinting at differential learning patterns between the Ability conditions for Self vs. Other.

Supplementary Note 3.2: Posterior predictive checks: Behavioral analyses on the predicted data.

To assess whether our winning model captured the core effects in our model free analysis, we let the parametrized winning model predict the time course of EXP for each participant, and compared these model predictions against the actual data (see Figure 3.1c). Figure 3.1c visually confirms the ability of the model to capture the observed data despite its small number of parameters. We repeated the behavioral analyses we had done on the actual behavioral data on the predicted data. The Trial x Ability condition x Agent condition linear mixed model on the predicted data revealed a significant main effect of Ability condition (β=-19.91, $t_{(68)}$=-14.55, p<.001, *95% CI*=[-22.59; -17.22]) and interaction of Trial x Ability condition (β=-1.29, $t_{(5240)}$=-55.98, p<.001, *95% CI*=[-1.34; -1.25]) replicating the effect that participants learned over time. More negative performance expectations for the self could also be replicated as indicated by the main effect of Agent (β=6.88, $t_{(68)}$=6.50, p<.001, *95% CI*=[4.80; 8.95]). The significant interaction of Agent condition x Ability condition (β=1.12, $t_{(5240)}$=4.22, p<.001, *95% CI*=[0.60; 1.65]) indicated differential ability beliefs between the Ability conditions for self vs other. Only the three-way interaction of Trial x Agent condition x Ability condition (β=0.04, $t_{(5240)}$=0.88, p=.377, *95% CI*=[-0.05; 0.13]) failed to reach significance. The analysis largely repeats the behavioral analysis done on the model-free data onto the predictions thus confirming that it recapitulates the main effects in our data.

Supplementary Note 3.3: Model parameter checks.

To assess whether introducing *w* as an additional parameter compared to our previous publications (Müller-Pinzler et al., 2019) we assessed correlations between the same parameters from the simple Valence Model 5 and our winning Model 8. Parameter correlations were rather high (Valence Learning Bias: ρ=.87, p<.001, $\alpha_{Self/PE+}$: ρ=.82, p<.001, $\alpha_{Self/PE-}$: ρ=.86, p<.001) and we also found a similar negative Valence Learning Bias in both models (testing both Valence Learning Biases against zero: Model 8: *mean*=-.12, *t*=-2.97, *p*=.004; Model 5: *mean*=-.12, *t*=-3.13 *p*=.003) indicating that including w does not change the results or interpretation in a meaningful way.

Supplementary Note 3.4: Neural activations associated with feedback processing indicate a specific role of Prediction Error Sign during self-related learning.

To examine the brain processes that underlie how people form self- and other-related ability beliefs, we compared neural activation during feedback processing as measured with fMRI. On the subject level, the fixed-effects GLM assessing effects of the different feedback conditions included four epoch regressors modeling the hemodynamic responses to the different cue conditions (Ability: High vs. Low × Agent: Self vs. Other),

weighted with the performance expectation ratings per trial as parametric modulator for each condition. Four regressors modeled the four feedback conditions (Prediction Error Sign: Positive vs. Negative × Agent: Self vs. Other), each weighted with the PE value for each trial. Here, in line with the behavioral learning model, Prediction Error Sign does not correspond to Ability condition but refers to the categorical distinction between Feedback with positive PE vs negative PEs. One regressor modeled the performance expectation rating phase. The estimation periods for Self and Other were modeled as two regressors, and emotion ratings phase and the instruction phase as separate regressors. Each of the regressors was modeled with the exact duration as presented during the experiment: The cue phase was modeled with a duration of 2.5 secs, the expectation rating phase according to individual reaction times with a mean of 4.26 sec (SD=1.04), the estimation phase with 10 secs, the feedback phase with 3 secs, and the emotion rating phase with 22.51 sec (SD=3.85). To account for noise due to head movement, six additional regressors modeling head movement parameters were introduced and a constant term was included for each of the two sessions. On the second level, beta images for the four feedback conditions were included in a flexible factorial design with two repeated measurement factors (Prediction Error Sign and Agent).

We found that the bilateral insula, anterior cingulate cortex, and thalamus (amongst others, see Supplementary Figure 3.4 and Supplementary Table 3.7) were activated significantly more strongly for self-related compared to other-related performance feedback (i.e. Agent effect). This finding of heightened activity in brain regions that have been linked to arousal, but also to self-agency, potentially reflects a difference in the subjective salience of self- vs. other-related information (Craig, 2009a; Späti et al., 2014; Sperduti et al., 2011). Compared to feedback for the Self, feedback for the Other resulted in stronger activation of the left and right middle temporal gyrus and precuneus/ middle cingulate gyrus (Supplementary Table 3.7).

Second, we compared self-related positive vs. negative feedback in order to examine how the valence of information affected neural processing (categorical Prediction Error Sign effect). We found significantly stronger activations of the left and right nucleus accumbens/ ventral striatum (NAcc/VS), bilateral angular gyrus, medial prefrontal cortex (mPFC), and precuneus/ posterior cingulate cortex (PCC) for positive Prediction Error Sign than for negative Prediction Error Sign (see Supplementary Table 3.7). This valence effect was unique for the processing of self-related information and did not emerge for other-related performance feedback (no significant clusters for the Prediction Error Sign effect for Other; $p < .001$). The opposite contrast, negative vs. positive Prediction Error Sign, yielded no significant activations, either for self-related or for other-related information. When testing the interaction of Agent x Prediction Error Sign, we found increased activation for self-related positive vs. negative feedback ([Self positive PE > Self negative PE] > [Other positive PE > Other negative PE]) in the angular gyrus (see Supplementary Table 3.7), and at a more lenient threshold also the bilateral NAcc/VS, the precuneus/ PCC, and precentral gyrus (cluster-wise FWE-corrected with $p < .05$ at a

cluster forming threshold of $p < .001$; see Supplementary Figure 3.4 and Supplementary Table 3.8).

Supplementary Note 3.5: *Specific associations of embarrassment and pride with neural activity in response to self-related PE valence*

To test whether embarrassment and pride had independent effects on neural activity in response to self-related PE valence within our predefined ROIs, we extracted parameter estimated for the effect of the parametric weights for PE valence for each whole ROI. Mean parameter estimates for the whole ROIs were then entered into regression models predicting the neural activity with both affect ratings simultaneously. We found independent effects of pride (β=0.36, $t_{(36)}$=2.63, p=.012) and embarrassment (β=-0.39, $t_{(36)}$=-2.82, p=.008; R^2=.33, $F_{(2,36)}$=8.94, p<.001) within the amygdala. We also found independent effects of pride (β=0.43, $t_{(36)}$=3.17, p=.003) and embarrassment (β=-0.36, $t_{(36)}$=-2.63, p=.013; R^2=.36, $F_{(2,36)}$=10.10, p<.001) within the dAI. For the vAI we found a significant effect of pride (β=0.38, $t_{(36)}$=2.61, p=.013) and a trend-wise effect of embarrassment (β=-0.29, $t_{(36)}$=-2.01, p=.052; R^2=.26, $F_{(2,36)}$=6.47, p=.004). Independent effects for pride (β=0.39, $t_{(36)}$=2.85, p=.007) and embarrassment (β=-0.40, $t_{(36)}$=-2.91, p=.006; R^2=.36, $F_{(2,36)}$=9.97, p<.001) were also present for the mPFC and we also found independent effects of pride (β=0.32, $t_{(36)}$=2.39, p=.022) and embarrassment (β=-0.46, $t_{(36)}$=-3.37, p=.002; R^2=.36, $F_{(2,36)}$=10.20, p<.001) for the VTA/ SN.

Supplementary Note 3.6: *Assessment of dependencies between PE surprise and PE valence for linear mixed model analyses*

The sizes of signed PEs (PE valence) and unsigned PEs (PE surprise) always matches, i.e. there are no instances, by definition, in which a large signed PE co-occurs with a small unsigned PE, or vice versa ("tiling"). We, therefore, conducted some control analyses to assess whether the fixed effects (FFX) estimates obtained by fitting our mixed models are unaffected by this.

We simulated 200 sets of new pupil dilation data for each subject for five different degrees of tiling of signed and unsigned PEs (see below). We then obtained FFX estimates from each of these 200 sets of simulated data to test whether the distributions of FFX estimates obtained by our mixed models were affected by tiling. For both data simulation and model estimation we used the same underlying model as used to analyze the empirical data reported in the manuscript. All simulations were based on the empirical data and the distributions of effect estimates (FFX and RFX) obtained from these data. This means that the RFX intercepts and noise component in the simulated data were generated by drawing from a normal distribution with zero mean and a standard deviation equal to the RFX intercepts and residuals, respectively, from our empirical models. We only changed the data for unsigned PEs before generating a new set of simulated data: precisely, to induce different degrees of tiling, we shuffled a subset of the empirical unsigned PEs on each simulation run. The subset was randomly selected from all 1440 data points on each simulation and different percentages of these data points were shuffled (0%, 25%, 50%, 75%, or 100%). On each simulation run, we created a single set of new regression weights

for each participant which was then combined with each of the five levels of tiling to compute new data based on the first-level regression formulas. A noise component was added. Afterwards, these simulated datasets were analyzed with the same mixed model used in the manuscript, to recover FFX estimates for the parameters underlying each set of simulated data.

First, we verified that the FFX recovered from simulated data were in a domain of parameter space that, as intended, coheres with our empirical FFX (see Supplementary Figure 3.7). T-tests showed no significant differences between empirical and recovered FFX (all $p>.07$, uncorrected). Moreover, no differences between levels of tiling were found for directed (all $ps>.154$) or absolute deviations (all $ps>.184$) of recovered from empirical FFX. Together, this demonstrates that recovered FFX were in realistic domains of parameter space, and independent of tiling levels. Second, we tested whether the FFX used for data simulation were positively correlated with the recovered FFX. To do so, we tested correlations between each set of recovered FFX (e.g., the intercepts) and the corresponding parameter underlying the simulated data (see Supplementary Figure 3.8). Here, all correlations were positive (all $ps<.032$, Holm-corrected), with the only exception being the FFX for unsigned PEs in a single level of shuffling (100%; completely even tiling), which just missed significance with $p=.053$. In a final step, we tested whether the degree of tiling affected the quality of FFX recovery. For each fixed effect (intercept, signed PE, Valence Learning Bias, unsigned PE, Valence Learning Bias * signed PE) we tested whether levels of tiling induced over- or underestimations of FFX (as compared to the parameters underlying the simulated data) by using the differences between recovered and underlying parameters as the dependent variable. No such effect was found for any of the five FFX (all $p>.153$; see Supplementary Figure 3.9). In a similar fashion, we tested whether the precision of FFX recovery was affected by levels of tiling by performing equivalent analyses of variance for the absolute differences between recovered and underlying parameters. Again, no such effect was found for any parameter (all $p>.194$; see Supplementary Figure 3.9). Together, our analyses demonstrate that differences in tiling of signed and unsigned PEs did not induce systematic biases during data simulation or the estimation of FFX from these simulated data.

Supplementary Tables

Supplementary Table 3.1. PSIS-LOO Scores

Model	PSIS-LOO	LOO-SE	LOO-Diff (SE-Diff)	% of $\hat{k} > 0.7$	No. Est. Parameters
Mean Model (M0)	-2644.4	319.7	1436.1 (142.8)	0.07	4
Self = Other					
Unity Model (M1)	-1801.3	396.5	593.0 (109.0)	0.47	5
Context Model (M2)	-1681.2	367.8	472.9 (80.6)	0.58	6
Valence Model (M3)	-1679.3	388.0	470.9 (93.9)	0.74	6
Self ≠ Other					
Unity Model (M4)	-1621.2	363.6	412.9 (75.5)	0.34	6
Context Model (M5)	-1599.9	372.6	391.6 (69.2)	1.43	8
Valence Model (M6)	-1346.4	333.6	138.1 (39.0)	0.53	8
ext. Valence Model (M7)	-1251.4	349.2	43.1 (16.7)	1.58	9
ext. Valence Model (M8)	-1208.3	357.7	-	1.39	9

Note. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (extended Valence Model M8) and standard errors of differences; percentage of $\hat{k}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al., 2016; No. Est. Parameters = number of estimated parameters in the model.

Supplementary Table 3.2. Initial and final ability beliefs during the LOOP task.

		Initial Belief		Final Belief	
		Mean	SD	Mean	SD
Self					
	High Ability	51.4	8.3	61.1	15.7
	Low Ability	53.3	8.3	30.7	13.8
Other					
	High Ability	57.1	6.3	69.9	9.3
	Low Ability	58.9	5.9	39.7	14.1

Note. Mean performance expectation ratings for the first trial (inital belief) and for the last trial of the experiment (final belief) for each of the four Ability conditions. SD = standard deviation.

Model Parameters

Model Parameters	SV1	SV2	SV3	SV4	αSelf$_{/PE+}$	αSelf$_{/PE-}$	αOther$_{/PE+}$	αOther$_{/PE-}$	BiasSelf	BiasOther
SV1		0.58*	0.33*	0.13	0.34*	-0.34*	0.06	0.1	0.46*	-0.01
SV2	0.58*		0.11	0.30*	0.12	-0.19	-0.16	-0.24*	0.23	0.16
SV3	0.33*	0.11		0.21	0.23	0.03	0.28*	0.32*	0.07	-0.11
SV4	0.13	0.30*	0.21		0.12	-0.27*	-0.18	-0.31*	0.40*	0.14
αSelf$_{/PE+}$	0.34*	0.12	0.23	0.12		0.11	0.45*	0.43*	0.52*	0.04
αSelf$_{/PE-}$	-0.34*	-0.19	0.03	-0.27*	0.11		0.17	0.19	-0.70*	-0.08
αOther$_{/PE+}$	0.06	-0.16	0.28*	-0.18	0.45*	0.17		0.80*	0.06	0.14
αOther$_{/PE-}$	0.10	-0.24*	0.32*	-0.31*	0.43*	0.19	0.80*		0.02	-0.40*
BiasSelf	0.46*	0.23	0.07	0.40*	0.52*	-0.70*	0.06	0.02		0.113
BiasOther	-0.01	0.16	-0.11	0.14	0.04	-0.08	0.14	-0.40*	0.11	

Supplementary Table 3.3. Pearson correlations for all model parameters of the winning Valence Model 8 as well as the Valence Learning Bias for Self and Other calculated from the learning rates as described in the methods section. * p<.05.

Spearman Correlations

	Valence Learning Bias	Embarrassment	Pride	Happiness	Tension
Valence Learning Bias		-0.24*	0.55**	0.23	-0.08
Embarrassment	-0.24*		-0.10	-0.07	0.53**
Pride	0.55**	-0.10		0.39**	0.20
Happiness	0.23	-0.07	0.39**		-0.07
Tension	-0.08	0.53**	0.20	-0.07	

Supplementary Table 3.4. Spearman correlations for all emotion ratings and the Valence Learning Bias. * p<.05; ** p<.01.

Supplementary Table 3.5. Activations Associated with PE Surprise

Contrasts/ Brain regions	Side	Cluster Size	MNI Coordinates			*T*	*p*
			x	y	z		
Self: PE Surprise							
Paracingulate Gyrus/ Superior Frontal Gyrus	R/L	12	-6	50	20	5.75	.009
Temporal Pole/ Frontal Orbital Cortex	L	8	-39	17	-22	5.57	.014
Superior Frontal Gyrus	R	1	12	20	62	5.20	.037
Temporal Pole/ Frontal Orbital Cortex	L	1	-30	11	-28	5.10	.048
Other: PE Surprise							
Temporal Pole/ Frontal Orbital Cortex	L	16	-33	17	-28	6.90	.001
Temporal Pole/ Frontal Orbital Cortex	R	25	39	20	-28	6.75	.001
Angular Gyrus/ Posterior Supramarginal Gyrus	R	23	48	-46	26	6.48	.002
Superior Frontal Gyrus/ Frontal Pole	R/L	74	6	53	26	6.41	.002
			-3	53	23	6.24	.003
Frontal Pole/ Superior Frontal Gyrus	R	3	9	47	47	5.50	.023
Posterior Supramarginal Gyrus/ Angular Gyrus	L	1	-51	-49	17	5.35	.033
Temporal Pole	R	1	48	17	-31	5.27	.042

Note. PE surprise refers to the unsigned prediction error values as parametric modulator for the feedback phase. The *p*-values are FWE corrected at peak level for the whole brain.

Supplementary Table 3.6. Activations Associated with PE Valence

Contrasts/ Brain Regions	Side	Cluster Size	MNI Coordinates			*T*	*p*
			x	y	z		
Self: PE Valence							
Superior/ Middle Frontal Gyrus	L	197	-15	29	53	7.07	<.001
Middle/ Superior Frontal Gyrus			-36	17	50	6.38	.002
Middle Frontal Gyrus			-39	23	38	5.75	.009
Superior Parietal Lobule/Superior Lateral Occipital Cortex	L	199	-36	-58	56	6.96	<.001
Angular Gyrus/ Posterior Supramarginal Gyrus			-45	-55	35	6.77	.001
Caudate / Accumbens	L	139	-9	20	-1	6.76	.001
Caudate / Accumbens	R		12	17	-1	6.41	.002
Superior Lateral Occipital Gyrus/ Angular Gyrus	R	121	48	-61	41	6.48	.001
Posterior Supramarginal Gyrus/ Angular Gyrus			51	-46	47	6.34	.002
Superior Parietal Lobule/ Angular Gyrus			39	-55	56	5.44	.020
Postcentral Gyrus/ Superior Parietal Lobule	L	50	-45	-34	56	6.08	.004
Postcentral Gyrus/ Posterior Supragarginal Gyrus			-45	-28	41	5.53	.016
Self > Other: PE Valence							
Accumbens	L	19	-9	26	-1	5.77	0.008
Caudate/ Accumbens	R	2	12	17	-4	5.23	0.034

Note. PE valence refers to the signed prediction error values as parametric modulator for the feedback phase. The *p*-values are FWE corrected for the whole brain at peak level. Only clusters with more than 50 voxels are reported for the Self: PE Valence contrast.

Supplementary Table 3.7. Activations Associated with Feedback Processing: Interaction of Agent * Prediction Error Sign

Contrasts/ Brain Regions	Side	Cluster Size	MNI Coordinates			*T*	*p*
			x	y	z		
Interaction: Self > Other, Positive > Negative							
Angular Gyrus/ Superior Lateral Occipital Cortex	R	229	48	-58	41	5.28	.002
			57	-61	20	3.55	
Angular Gyrus / Superior Parietal Lobule	L	303	-42	-55	44	4.78	.001
			-42	-55	53	4.76	
Angular Gurus / Posterior Supramarginal Gyrus			-48	-55	29	4.57	
Putamen/ Pallidum	R	380	18	5	-10	4.67	<.001
Caudate / Accumbens	R		12	20	-1	4.56	
Caudate / Accumbens	L		-9	20	-1	4.55	
Precentral gyrus	L	162	-18	-19	53	3.95	.008
			-27	-13	44	3.79	
			-24	-25	41	3.71	
Posterior Cingulate Gyrus/ Precuneous Cortex	R/L	296	-15	-43	32	3.91	.001
Precuneous Cortex/ Posterior Cingulate Gyrus			-3	-58	35	3.87	
Posterior Cingulate Gyrus/ Precuneous Cortex			6	-43	26	3.73	
Cerebellum Left Crus I / Crus II	L	154	-12	-82	-25	3.73	.009
Occipital Fusiform Gyrus / Cerebellum Left Crus I			-30	-79	-1	3.69	
Occipital Fusiform Gyrus / Inferior Lateral Occipital Cortex			-30	-85	-10	3.65	

Note. Cluster extents refer to $p < .001$, uncorrected and *p*-values are FWE corrected on the cluster level.

Supplementary Table 3.8. Differential Functional Connectivity of the Dorsal Anterior Insula Associated with PE Valence

Covariates/ Regions of interest	Side	Cluster Size	MNI Coordinates			*T*	*p*
			x	y	z		
PPI Right Dorsal Anterior Insula							
Amygdala	R	6	33	-1	-31	4.46	.003
	L	2	-27	-4	-25	3.89	.013
		5	-24	2	-13	3.68	.022
Ventral Tegmental Area/ Substantia Nigra	R/L	5	-18	-16	-13	4.07	.015
Medial Prefrontal Cortex	R/L	11	-9	35	53	4.95	.005
		5	-6	59	26	4.62	.012
PPI Left Dorsal Anterior Insula							
Amygdala	L	3	-30	-4	-22	3.76	.019
Ventral Tegmental Area/ Substantia Nigra	R/L	3	-9	-13	-13	3.66	.042

Note. Stronger functional connectivity for negative vs positive PEs for self- vs other-related feedback. The *p*-values are FWE corrected within ROIs at peak level.

Supplementary Table 3.9. Sample Characteristics

	fMRI Sample		Behavioral Sample		p
	Mean	SD	Mean	SD	
Age	22.30	2.65	23.30	3.97	.234
Self-esteem	6.24	0.84	6.07	1.26	.522

Note. Sample characteristics for both samples. SD = standard deviation; fMRI Sample: n=39, Behavioral Sample: n = 30; p-value refers to a two-sample t-test, df = 67.

Supplementary Data

Supplementary Data 1 – 3 can be found under the following link:

Neurocomputational mechanisms of affected beliefs | Communications Biology

(https://www.nature.com/articles/s42003-022-04165-3#Sec26)

Supplementary References

Craig, A. D. B. (2009). How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*(1), 59–70. https://doi.org/10.1038/nrn2555

Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-50821-w

Späti, J., Chumbley, J., Brakowski, J., Dörig, N., Grosse Holtforth, M., Seifritz, E., & Spinelli, S. (2014). Functional lateralization of the anterior insula during feedback processing. *Human Brain Mapping*, *35*(9), 4428–4439. https://doi.org/10.1002/hbm.22484

Sperduti, M., Delaveau, P., Fossati, P., & Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: A brief review and meta-analysis. *Brain Structure and Function*, *216*(2), 151–157. https://doi.org/10.1007/s00429-010-0298-1

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, *17*(1), 3581–3618

4 Study 3

Neurocomputational Mechanisms Underlying Maladaptive Self-Belief Formation in Depression[3]

4.1 Abstract

Maladaptive self-beliefs are a core symptom of major depressive disorder. These are perpetuated by negatively biased feedback processing. Understanding the neurocomputational mechanisms of biased belief updating may help to counteract maladaptive beliefs. The present study uses functional neuroimaging to examine neural activity associated with prediction error-based learning in persons with major depression and healthy controls. We hypothesized that increased symptom burden is associated with negatively biased self-belief formation and altered neural tracking of social feedback. Results showed that a higher symptom burden was associated with forming more negative self-beliefs and more positive beliefs about others. This bias was driven by reduced learning from positive prediction errors in depression. Neural reactivity of the insula showed increased tracking of more negative self-related prediction errors. The interplay of increased neural responsiveness to negative feedback and reduced learning from positive feedback may contribute to the persistence of maladaptive self-beliefs and, thus, the maintenance of depression.

---

[3] This study has been uploaded on a preprint server as: **Czekalla, N.**, Schröder, A., Mayer, A. V., Stierand, J., Stolz, D. S., Kube, T., Wilhelm-Groch, I., Klein, J. P., Paulus, F. M., Krach, S., & Müller-Pinzler, L. (2024). *Neurocomputational Mechanisms Underlying Maladaptive Self-Belief Formation in Depression.* In bioRxiv (p. 2024.05.09.593087). https://doi.org/10.1101/2024.05.09.593087. Currently in the submission process.
My contribution: designing the research, data acquisition, data analysis, and writing the manuscript.

4.2 Introduction

Maladaptive, mostly negatively biased, and highly rigid self-beliefs are prominent in various mental disorders (Beck, 1976). Humans act on their beliefs as if they reflect reality (Bandura, 1977), which makes maladaptive beliefs a key maintaining factor (Beck, 1979). Self-beliefs and beliefs about the world are a product of our learning history and are constantly updated in the face of incoming information that deviates from expectations (Markus & Wurf, 1987), also called prediction errors (Friston, 2005; Rouhani et al., 2023). In the case of depression, maladaptive self-beliefs often revolve around low self-efficacy, for example, "I can't make it anyway" (Bandura et al., 1999; Hollon & Kendall, 1980). When the emphasis on this internal model is strong (Clark et al., 2018), one becomes insensitive to the context and may neglect contradictory positive feedback (Kube et al., 2020; Villano & Heller, 2024). To understand the mechanisms of this maladaptive self-belief formation, we assessed the processes underlying ongoing self-belief formation in a computational modeling approach (Müller-Pinzler et al., 2019, 2022) in individuals diagnosed with depression and healthy control participants in a functional magnetic resonance imaging (fMRI) study.

Negative beliefs about oneself, the world, and the future are typical characteristics of depression, which are maintained by negatively distorted information processing (Beck, 1963, 1964, 1979). Accordingly, a more negatively biased updating of beliefs about one's future, abilities, or social popularity has been observed in experimental settings in depression (Garrett et al., 2014; Korn et al., 2014) or in association with depression-related symptoms like low self-esteem (Müller-Pinzler et al., 2019; Will et al., 2020). More recently, mental disorders have been conceptualized as an interplay of maladaptive prior internal models with a maladaptive response to prediction errors (Clark et al., 2018; Stephan et al., 2016; Sterzer et al., 2018), which is the mismatch between prediction and actual outcome. The persistence of negative beliefs, in particular in depression, has been related to reduced updating after unexpected positive information (Kube et al., 2020). This has been shown, for example, after receiving self-related positive performance feedback (Kube et al., 2019) or unexpectedly positive social reactions in imagined situations (Everaert et al., 2018). A potential cognitive mechanism that might hinder the change of negative beliefs in the face of positively disconfirming information is "cognitive immunization" (Kube et al., 2019). This reflects the reappraisal of disconfirming feedback, such that feedback is devalued and people's initial expectations are maintained. For example, people with depression could discount positive feedback by viewing it as an exception (e.g., "I was just lucky") or doubting its credibility (e.g., "She was only nice to me because she wants my help"; Kube et al., 2019; Rief et al., 2015). In cognitive behavioral therapy, this process is referred to as cognitive distortion, "discounting the positive (Dobson & Dozois, 2021)."

The experience of prediction errors is also an affective phenomenon (Barrett, 2017; Eldar et al., 2016), which is of particular importance in the context of belief updating in depression. Negative affect is one of the central depressive symptoms (American Psychiatric Association, 2022), and the individual's current affective state shapes

expectations (Bennett et al., 2023), which thereby contribute to the formation of self-beliefs (Bromberg-Martin & Sharot, 2020). This, in turn, shapes how the individual feels in a particular situation, which results in a recursive influence of affect and belief updating on each other (Bromberg-Martin & Sharot, 2020; Eldar & Niv, 2015). In this line, affective experiences during learning have been linked to biases in forming self-beliefs and neural processes (Müller-Pinzler et al., 2022). In this sense, individuals show less flexibility in adjusting their beliefs in a positive direction when in a low mood (Karnick et al., 2024; Kube et al., 2023). Accordingly, the effectiveness of psychological interventions can increase if, in addition to challenging maladaptive beliefs on a cognitive level, emotional meanings and in-session emotional experiences are also addressed (Samoilov & Goldfried, 2000). This illustrates the importance of the interplay of affect and cognitive strategies for changing maladaptive self-beliefs.

The entanglement of forming self-beliefs and affective experiences also manifests on the neural systems level (Müller-Pinzler et al., 2022). Several studies in healthy samples could show that valence-dependent tracking of prediction errors in the ventral striatum (VS) was more robust for self-related than other-related information (Müller-Pinzler et al., 2022) and that the neural prediction error signal is associated with biases in updating self-beliefs (Kuzmanovic et al., 2016; Sharot et al., 2011). Dopaminergic regions such as the VS and also midbrain nuclei in the ventral tegmental area and substantia nigra (VTA/SN) are well described for coding prediction error signals (Diederen et al., 2016; Schultz, 1998). However, regions linked to affective experience and motivational processes are also involved in prediction error processing. The insula, involved in emotional processing and body sensations as well as attention and action monitoring (Koban & Pourtois, 2014; Müller-Pinzler et al., 2015; Touroutoglou et al., 2012), shows modulated neural prediction error signaling, especially if they convey negative information (Kumar et al., 2018; Mulej Bratec et al., 2015; Rothkirch et al., 2017; Seymour et al., 2005; Waltz et al., 2009). Also, the amygdala, which is linked to emotional learning (Murray, 2007; Phelps, 2006), is involved in tracking prediction errors (Kumar et al., 2008; McHugh et al., 2014; Seymour et al., 2005). Accordingly, the neural prediction error signals for self-related information in the anterior insula, the amygdala, and in the VTA/SN were associated not only with biased processes of self-belief formation but also with current affective states, which led to the idea of “affected beliefs” (Müller-Pinzler et al., 2022).

In the context of depression, most studies focused on the neural processing of reward prediction errors and behavioral adaption following rewarding information unrelated to a self-belief (i.e., reward learning, Dayan & Niv, 2008). In line with the persistence of negative beliefs against conflicting positive information, many studies reported reduced reward learning in depression (Admon & Pizzagalli, 2015; Kumar et al., 2018; Robinson et al., 2012; Safra et al., 2019). However, other studies found no support for different reward learning (Brolsma et al., 2022; Gradin et al., 2011; Rothkirch et al., 2017; Rouhani & Niv, 2019) or prediction error-dependent fluctuation in state happiness between individuals with and without depression, indicating intact reward processing (Rutledge et al., 2017). At the neural systems level, studies found reduced reward-related prediction

error signaling in the ventral striatum (Gradin et al., 2011; Kumar et al., 2008, 2018; Robinson et al., 2012) and other brain regions, such as the ventral tegmental area (Kumar et al., 2018), anterior cingulate cortex, and hippocampus (Chen et al., 2015; Gradin et al., 2011; Kumar et al., 2008). However, effects were also mixed, and other studies reported unaltered reward prediction error signaling in depression (Rothkirch et al., 2017; Rutledge et al., 2017). In response to negative prediction errors, some studies reported increased activity (Ubl et al., 2014), while others did not (Kumar et al., 2018; Rothkirch et al., 2017).

Concerning more specific updating of self-related beliefs regarding one's future, depression was associated with greater tracking of negative prediction errors in the right inferior parietal lobule and inferior frontal gyrus (Garrett et al., 2014). In individuals with social anxiety, which is also characterized by negative self-beliefs (Heimberg et al., 2014) and often comorbid with depression (Adams et al., 2016; Kessler et al., 1999), activity in the insula mediated the effect of negative social feedback on self-belief updates (Koban et al., 2023). Hypersensitivity to negative information in the insula has also been reported for depression (Engelmann et al., 2017) and has been linked to the way emotions are processed in this condition (Mutschler et al., 2012; Sliz & Hayley, 2012). Moreover, the induction of negative affect increased reactivity to negative information in the insula (Harlé et al., 2012).

While prediction errors are relevant for changing beliefs and learning from prediction errors has been widely studied in depression, the link to the formation of self-beliefs is still missing. The current study builds on previous work by applying a well-established computational approach of trial-by-trial self-belief formation, the Learning Of Own Performance (LOOP) task (Czekalla et al., 2021; Müller-Pinzler et al., 2015, 2019, 2022), to a clinical sample with depression. To understand the mechanisms of how people arrive at their maladaptive self-beliefs, we aimed to test whether 1) depression is related to biased updating of self-beliefs, as opposed to updating beliefs about another person, and 2) whether this is underpinned by altered neural prediction error processing in individuals diagnosed with depression. To test these hypotheses, we followed a two-step procedure for our analyses. After a group comparison approach with participants diagnosed with depression vs. healthy controls, we followed a transdiagnostic dimensional approach following the recommendations of the Research Domain Criteria guidelines (Brolsma et al., 2022; Insel et al., 2010). In line with negative self-beliefs and distorted information processing in depression, we expected more negatively biased processes of belief formation, specifically if the information is related to the self. As we previously showed associations between prediction error processing and biased self-belief formation and affect in the insula, VTA/SN, and amygdala using the same paradigm (Müller-Pinzler et al., 2022), we expected depression-related alterations in these regions. In line with other studies on reward learning and depression, we included the ventral striatum as a region of interest (ROI). The results show that a higher symptom burden is associated with forming more negative self-beliefs and more positive beliefs about others. Neural activity

of the insula showed increased tracking of more negative self-related prediction errors but no difference in tracking positive prediction errors.

4.3 Results

Measuring neural processes of belief formation

Participants diagnosed with depression (*n*=35) and healthy control participants (*n*=32) completed the Learning Of Own Performance (LOOP) task (Müller-Pinzler et al., 2019, 2022) in the MRI (for sample characteristics, see Supplementary Table 4.1). In the LOOP task, participants were asked to estimate specific attributes of objects (e.g., the height of buildings or the weight of animals). By incorporating manipulated performance feedback, participants were led to form novel beliefs about their own (Self) or another person's (Other) estimation abilities. The other person allegedly performed the task simultaneously in an adjacent room outside the MRI. At the beginning of each trial, a screen displayed the estimation category of the upcoming trial. It indicated whether participants had to perform a Self-trial or an Other-trial, followed by a rating of expected performance. On a trial-by-trial basis, participants received manipulated performance feedback that was more positive in one condition and negative in the other, allowing for an assessment of learning biases during belief formation (Figure 4.1A). This resulted in four feedback conditions: Agent (Self vs. Other) x Ability (High Ability vs. Low Ability, Figure 4.1B).

As in previous studies, participants adjusted their expected performance ratings according to the feedback over time, i.e., they formed novel beliefs about their own and another person's ability (Figure 4.1C, Supplementary Note 4.1, Supplementary Table 4.2). To describe the belief formation process, we modeled the changes in participants' expected performance through updates from prediction errors (Lockwood & Klein-Flügge, 2020). The central variable of interest here is the learning rate, i.e., the extent to which prediction errors are weighted to make an update. Consistent with our previous studies, the winning model included distinct learning rates for positive (LR+) and negative (LR-) prediction errors, separately for Self and Other, allowing a valence- and agent-specific description of belief updating (factors Prediction error valence [PE Val] and Agent; for a more detailed description of this model and the whole model space, see Methods section and Supplementary Table 4.3. For model comparison results in the total sample and the subsamples, see Supplementary Note 4.2 and Supplementary Table 4.4; for correlations of model parameters, Supplementary Figure 4.10). This model was selected for further analysis. To measure biased learning, we aggregated the learning rates into a Valence Bias Score VBS = (LR+ - LR-)/ (LR+ + LR-), separately for Self and Other, as in previous studies (Müller-Pinzler et al., 2019, 2022). We assessed self-reported symptom burden with a general measure of depressive symptoms (Schmitt et al., 2003) as well as a more specific measure of the cognitive symptoms of depression (Pössel et al., 2005). To address symptom burden in a broader context, we additionally assessed symptoms of social anxiety (Stangier et al., 1999) and self-esteem (Marsh & O'neill, 1984) as transdiagnostic markers, already proven to bias self-belief formation behavior (Müller-Pinzler et al., 2019). Variability in the severity of self-reported symptoms was well

explained by a single principle component (83% variance explained) reflecting general depressive and socially anxious psychopathology, akin to other recent advances (Hoven et al., 2023; Seow et al., 2021; Will et al., 2017), see 4.5 Methods section and Figure 4.1E).

Negativity bias in self-belief formation in individuals with and without depression

Both groups showed a negativity bias in forming novel self-beliefs, which replicated previous studies using this task (Czekalla et al., 2021; Müller-Pinzler et al., 2019, 2022). This bias means that learning rates were lower with positive than negative prediction errors. It was not present when participants formed beliefs about the other person's ability (PE_Val x Agent interaction: $t(195)=-3.29$, $p<.001$, negativity bias: post-hoc $LR[-]_{Self}$ vs $LR[+]_{Self}$; $t(66)=4.78$, $p<.001$). Between groups, there was no difference in the overall extent of learning (main effect Group $t(65)=-.30$, $p=.764$) and no difference in the extent of the negativity bias (PE_Val x Agent x Group interaction: $t(195)=.18$, $p=.861$; Figure 4.1C and D, Supplementary Table 4.5). This suggests that people diagnosed with depression did not differ from healthy people when forming novel beliefs in response to the feedback received. The negativity bias in both groups, which may be due to rather negative prior beliefs regarding their estimation abilities and negative affect elicited by the performance context of the setting (Müller-Pinzler et al., 2022), demonstrates the susceptibility to biases in self-related information processing.

Biased belief updating is linked to symptom burden

When addressing biased belief formation and symptom burden from a dimensional perspective, we found that learning biases were associated with self-reported symptoms. With higher scores in psychopathology, participants learned more negatively about themselves (VBS_{Self}) but more positively about others (VBS_{Other}, Agent x Psychopathology interaction: $t(65)=-2.03$, $p=.047$, Figure 4.1F, Supplementary Table 4.6). This relative devaluation of one's performance in comparison to another person has been described to play an essential role in both the etiology and maintenance of depression (McCarthy & Morina, 2020; Swallow & Kuiper, 1988; Weary et al., 1987).

In the absence of group differences during belief formation, this effect suggests that within-group variance is particularly important. While many studies of depression only examine subclinical depression in healthy samples, we were able to look at the distribution of symptom burden in both groups. Our findings suggest that the association of negatively biased processes of belief formation with increased symptom burden was mainly driven by the clinical sample (VBS_{Self}: $r_{MDD}=-.43$, $r_{CON}=.05$, difference $z=-1.99$, $p=.047$; Supplementary Note 4.4, Supplementary Figure 4.1). This highlights the importance of including individuals with more pronounced symptoms when examining biases in belief formation in depression.

Disregard of positive information in association with symptom burden

Within the clinical sample, symptom burden was more strongly linked to a lower learning rate for positive prediction errors than to a higher learning rate for negative prediction errors (MDD: $rLR[+]_{Self}=-.51$, $rLR[-]_{Self}=.03$; difference of absolute values: $z=2.1$, $p=.036$,

Figure 4.1G, Supplementary Figure 4.2, for more detailed correlations with symptoms scores, see Supplementary Figure 4.3). This implies that people with more severe symptoms made fewer belief updates after positive prediction errors, which may drive the association of symptom burden and biased learning.

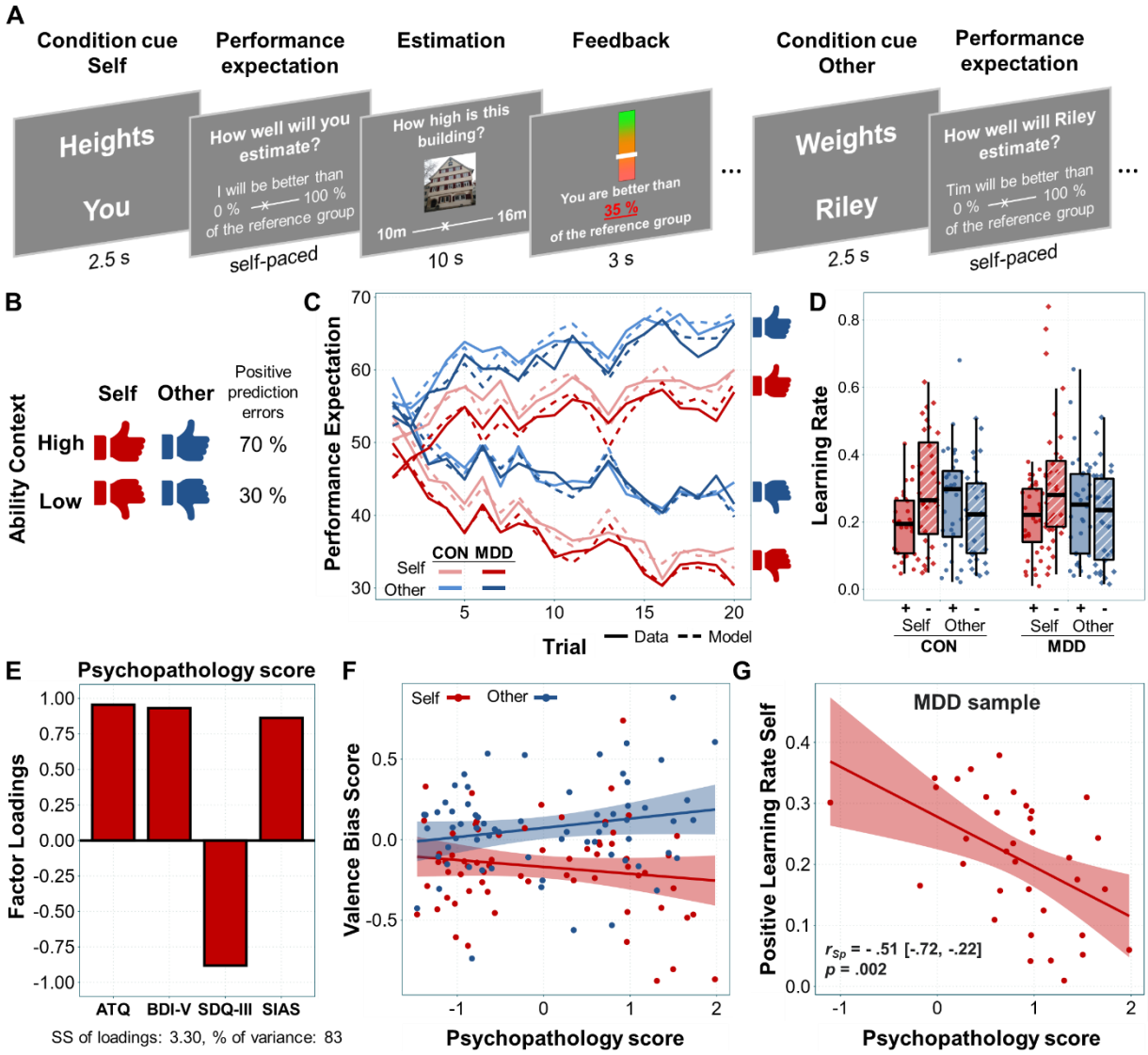


Figure 4.1. Trial sequence of the Learning Of Own Performance task, experimental conditions, and behavioral results. **A** Sequence of One Trial. 1.) Condition Cue: estimation category (e.g., weights) paired with either high or low ability feedback and agent, 2.) Rating of expected performance, 3.) Estimation question, 4.) Feedback. **B** Low and High Ability conditions of alleged estimation performance for Self and Other (Agent condition) were presented in pseudo-randomized order. Feedback with controlled prediction errors: High Ability: 70 %, Low Ability: 30 % of the trials with planned positive prediction errors (the actual percentage of positive/ negative prediction errors could slightly differ, e.g., if the feedback had been out of range). **C** Predicted and actual ratings of expected performance for both groups over time. The ratings indicate that participants update their expected performance (solid lines) according to the feedback, thus forming a belief about the performance levels in the different estimation categories. The winning model's predicted values (dashed lines) with separate learning rates for Agent and Prediction error valence captured the participants' behavior, as indicated by a close match of actual ratings and predicted values. **D** Learning rates of the winning model show a bias towards increased updating in response to negative prediction errors (-) in contrast to positive prediction errors (+), specifically for self-beliefs in both groups. Colored bars indicate the first and third quartile of the data; the line marks the median. Whiskers extend from the upper/ lower box borders to the largest/ smallest

data point at most 1.5 times the interquartile range above/ below the respective border. MDD=Major Depressive Disorder (n=35), CON=control group (n=32). **E** Factor loadings of the principal component analysis, by which one factor, a summary measure of symptom burden, was extracted from the questionnaire variables (Automatic Thought Questionnaire, ATQ; Pössel et al., 2005; Beck's depression inventory, BDI-V; Schmitt et al., 2003; Self-Description Questionnaire-III, SDQ-III subscale scores; Marsh & O'neill, 1984; Social Interaction Anxiety scale, SIAS; Stangier et al., 1999). SS = sum of squares. **F** Valence Bias Score (LR[+] – LR[-])/ (LR[+] + LR[-]) as a function of the Psychopathology score and Agent. A significant Agent x Psychopathology interaction shows more negatively biased updating for Self and more positively biased updating for Other with more symptom burden. Lines for displaying purposes. **G** Correlation of the positive learning rate (LR[+]) and the Psychopathology score in the clinical sample. Spearman correlation coefficient, line for displaying purposes.

Neural tracking of negative prediction errors in depression and association with affective experience

We examined the neural processing of self-related prediction errors to further understand the biased formation of self-beliefs. Individuals who updated their beliefs more positively (positive learning rate; LR[+]) also had stronger activity in the VTA/SN ROI after experiencing positive prediction errors (categorical PE_Val effect, Spearman correlation between averaged parameter estimates within ROI r=.33, p=.006). Individuals who updated their beliefs more negatively (negative learning rate; LR[-]) had stronger activity in the VTA/SN, insula, and amygdala ROIs during negative prediction errors (Spearman correlations within ROIs $rs \geq .27$, $ps \leq .029$, see Supplementary Figure 4.5). This demonstrates the link of neural prediction error signals and behavioral updating of self-beliefs.

When comparing individuals with depression and healthy controls, a significant Group x PE Valence interaction in the right ventral and posterior insula indicated an imbalance of negative relative to positive prediction errors signaling in the clinical sample, but not in healthy controls (continuous PE effect, $ps \leq .030$ family-wise [FWE] corrected at peak level within ROIs, Figure 4.2, Supplementary Table 4.8, for baseline activations of continuous PE effect see Supplementary Table 4.7). We found no difference between study groups in the association of neural activity with more positive self-related prediction errors (with $p<0.05$ FWE-corrected at peak level for all ROIs). This suggests that during the formation of novel self-beliefs, surprisingly positive feedback is processed similarly in individuals with and without depression. However, compared to control participants, individuals with depression showed stronger activity in the right dorsal, ventral, and posterior insula as well as in the VTA/SN with more negative prediction errors (two-sample t-tests, $ps \leq .046$ within ROIs, see Supplementary Table 4.8). This is in line with a generally stronger tracking of negative compared to positive prediction errors in the insula (Supplementary Figure 4.6) and suggests that individuals with depression are more sensitive to negative prediction errors at the neural level. In contrast, the processing of positive prediction errors is unaltered.

In addition, stronger activity in the same regions (right dorsal, ventral, posterior insula) with more negative prediction errors was accompanied by decreased happiness (Spearman correlation between averaged parameter estimates within all three ROIs *rs* ≥ -0.33, *ps* ≤ .029; Supplementary Figure 4.7). This finding of reduced positive affect during the processing of negative prediction errors is of particular importance in the context of affective disorders, as it is included in various psychological models of depression as a sustaining element in the vicious circle (Gross & Muñoz, 1995; Lewinsohn, 1974; Maier & Seligman, 1976; Smith et al., 2018). Stronger neural responsivity to more negative prediction errors and more negative affective responses could, therefore, influence each other and maintain the generally lower positive affect in individuals with depression during task performance (Supplementary Table 4.9).

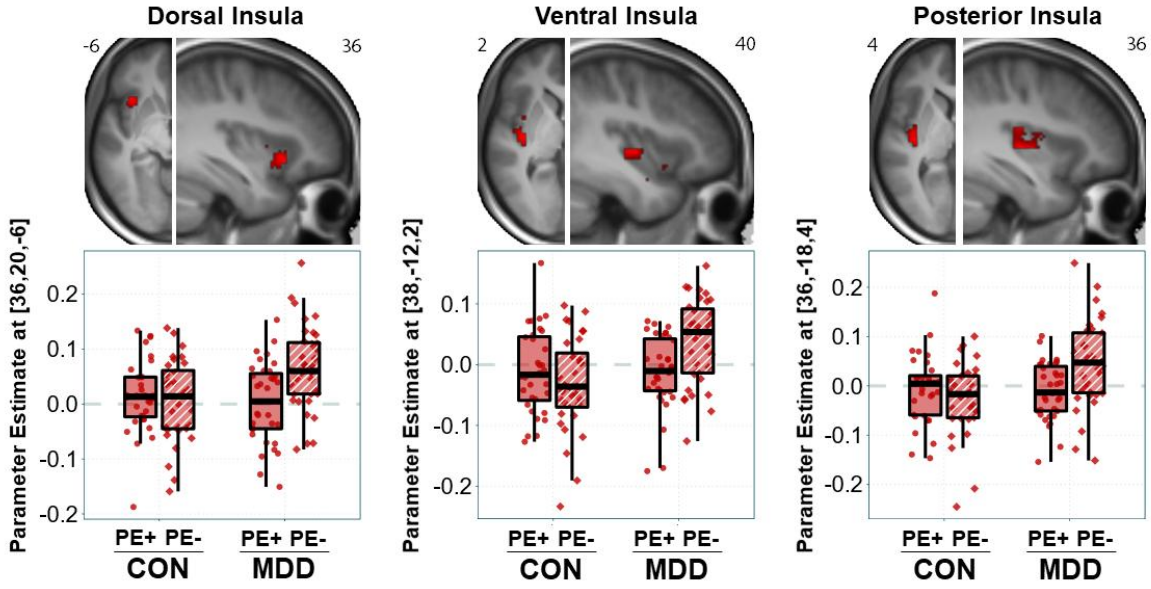


Figure 4.2. Differential insula tracking of prediction error valence in depression. The interaction (MDD[PE-]>MDD [PE+])>(CON [PE-]>CON [PE-]) shows a greater difference between PE- vs. PE+ in the MDD compared to the CON group (dorsal: *t*=2.90, *p*=.070, ventral: *t*=3.35, *p*=.029, posterior: *t*=3.17, *p*=.030, FWE corrected within ROIs at peak level). CON=control group (n=32), MDD=Major Depressive Disorder (n=35). The plot depicts the parameter estimates of the Group [CON, MDD] x PE Valence [PE+, PE-] interaction at peak level for the three regions of interest in the right insula. Brain plot: uncorrected *p*<0.05 within ROI for display purposes; see Supplementary Table 4.8 for FWE corrected statistics.

Neural tracking of negative prediction errors is linked to biased belief formation and symptom burden

In line with the results from the group comparison, symptom burden was associated with stronger responses to more negative prediction errors within the right dorsal and posterior insula (Spearman correlation with Psychopathology score within ROIs *rs*≥.26, *ps*≤.037). Stronger responses in the right dorsal insula with more negative prediction errors were also associated with a more negatively biased self-belief formation (VBS_{Self} *r*=-.24, *p*=.046), replicating previous findings (Müller-Pinzler et al., 2022). Neither levels of symptom burden nor the Valence Bias Score were significantly associated with neural activity in any of the ROIs with more positive prediction errors (Supplementary Figure 4.9).

These results suggest that a stronger tracking of negative prediction errors in parts of the insula accompanies more biased belief updating and higher symptom burden.

4.4 Discussion

In the present study, we used neurocomputational methods to examine the mechanisms that underlie the altered processing of self-related feedback in the formation of maladaptive self-beliefs in depression. To promote a transdiagnostic dimensional approach, we recruited a naturalistic clinical sample of individuals diagnosed with depression, including comorbidities of social anxiety, and examined the symptom burden dimensionally. Our data suggest that the combination of stronger neural reactivity of the insula during self-related negative prediction errors, together with greater neglect of self-related positive prediction errors during the updating process, might play a key role in the maintenance of maladaptive self-beliefs in depression. Our results provide new insights into cognitive distortions and the persistence of negative self-beliefs as important characteristics of the psychopathology of depression.

Behaviorally, the results replicate previous findings of a negativity bias in the formation of novel self-beliefs (Czekalla et al., 2021; Müller-Pinzler et al., 2019, 2022), both for individuals with depression and healthy controls. This bias is absent when participants form beliefs about others, highlighting the specificity of the valence bias for processing self-related information (Müller-Pinzler et al., 2022). While other studies report a positivity bias when updating self-beliefs about one's future (Sharot et al., 2011; Sharot & Garrett, 2016; Vandendriessche & Palminteri, 2023), personality (Korn et al., 2012), intelligence, or appearance (Eil & Rao, 2010; Villano et al., 2023), we find a negativity bias, which aligns with other studies using task-specific performance feedback (Brotzeller & Gollwitzer, 2024; Ertac, 2011; Zamfir & Dayan, 2022). This suggests that contextual effects, such as whether study participants learn about an ability vs. their personality, might impact the direction of learning biases. Besides the effects of the context, prior beliefs and confidence might be linked to the formation and revision of self-beliefs.

Although the dimensional approach suggests an association between depressive symptoms and biased belief formation, the simple group comparison demonstrates that individuals with depression are not fundamentally different from control participants in terms of how they form novel self-beliefs in this performance context. While some studies report less positive self-belief updating in samples with depression (Korn et al., 2014; Kube et al., 2019), other studies, particularly those that employ reward learning tasks, also do not find group differences in reward learning rates (Brolsma et al., 2022; Rothkirch et al., 2017; Rouhani & Niv, 2019). In addition to contextual effects like task selection, the sample's composition may be an important factor related to the extent of learning bias and differences between groups. The present findings indicate that, even in a depressive state, people can acquire novel beliefs about their abilities in a way that does not categorically distinguish them from people without this diagnosis.

In the dimensional approach, the results indicated that individuals with higher depressive and social anxiety-related symptom burden exhibited more negatively biased self-belief formation. In contrast, they formed their beliefs about others in a positively biased manner.

This aligns with other studies linking biased self-belief updating to symptom burden (Engelmann et al., 2017; Everaert et al., 2018, 2020; Korn et al., 2014; Safra et al., 2019). The diminished value one places on one's accomplishments relative to others, along with the affective response to this self-esteem-debilitating social comparison, has been identified as a symptom-maintaining factor in depression (McCarthy & Morina, 2020).

In the absence of group differences, this effect highlights the importance of considering within-group variance in symptom burden. Diagnostic categories summarize a heterogeneous symptom constellation and have been criticized as non-distinct and not reflecting the continuous nature from at-risk through mild to severe symptoms (Hyman, 2021; Roefs et al., 2022). Furthermore, comorbidities are common (Kessler et al., 2005). This means that heterogeneity in symptoms must be considered in clinical samples, especially in naturalistic samples, including comorbidities, as in the current study. A dimensional approach with a naturalistic sample considers the continuous nature of psychopathology and increases ecological validity (Morrison et al., 2003). The results indicate that the correlation between symptom burden and negatively biased self-belief formation was mainly driven by the clinical sample. Therefore, it is necessary to include individuals with clinically relevant symptom burden. Also, within the facets of psychopathology, those targeting specific symptoms such as negative thoughts or self-esteem were more likely to yield associations with biased learning, which may provide evidence for focusing on symptoms rather than categories.

In the clinical sample, the bias in belief formation is driven by reduced learning from positive self-related prediction errors. One explanation for this finding can be that reduced attention is allocated to positive feedback (Joormann & Quinn, 2014; Sears et al., 2011). An alternative explanation is that cognitive mechanisms are actively involved in devaluing positive feedback, resulting in a cognitive immunization against it (Kube, 2023). Future studies may employ experimental manipulations to modulate feedback value and attentional allocation as well as measure the evaluation of feedback and the attentional focus (potentially using eye-tracking) to further disentangle these possibilities. Addressing these aspects will provide a more nuanced understanding of the underlying cognitive processes contributing to the observed bias in self-belief formation. Overall, these results suggest that sensitivity to context decreases with increasing symptom burden, possibly due to an overemphasis on an already established negative internal model. This means that less is learned from positive prediction errors, and beliefs are biased towards prior belief models in a confirmatory way (Klayman, 1995; Palminteri & Lebreton, 2022).

This finding may have important clinical implications for psychotherapy. Rather than directly challenging the negative self-concepts of patients with depression (e.g., by cognitive restructuring), they can be supported in becoming more context-sensitive so that positive feedback that contradicts the internal model can be incorporated. This might be done by practicing skills that help to become less reactive to established negative self-beliefs, e.g., defusion in Acceptance and Commitment Therapy (ACT; Hayes et al., 1999) or detached mindfulness in Metacognitive Therapy (Wells, 2011) and more sensitive to the present context, e.g., mindfulness skills in ACT or mindfulness-based cognitive

therapy (Segal et al., 2018). To update a maladaptive belief, it may be necessary for a therapist to thoroughly examine the meaning attributed to new experiences and identify potential cognitive distortions, such as retrospectively devaluing positive experiences. These mechanisms can be discussed with patients as a factor that maintains a negative self-image. Second, patients may be empowered that meeting the diagnostic criteria for a particular disorder does not necessarily mean that they are fundamentally different in their ability to form new self-beliefs. This can be important, for example, when introducing new antidepressant activities where new self-beliefs need to be built.

On the neural system level, we found that the insula is associated with stronger tracking of negative prediction errors in the clinical sample compared to the control group. There was no group difference for positive prediction errors. Tracking of negative prediction errors was also associated with symptom burden. The stronger neural response to negative prediction errors is in line with a general neural sensitivity to negative (social) prediction errors in the insula (Garrison et al., 2013; Müller-Pinzler et al., 2022; Wächter et al., 2009), as well as a stronger insula reactivity specifically in samples with depression (Engelmann et al., 2017) or social anxiety (Heitmann et al., 2014; Koban et al., 2023). This effect could be explained by an attentional bias in individuals with depression (Donaldson et al., 2007; Joormann & Quinn, 2014) in an affect-congruent way (Bennett et al., 2023), with either a faster orientation (Sears et al., 2011) or longer maintenance of attention to negative feedback (Eizenman et al., 2003; Koster et al., 2005). Since the insula reactivity to more negative prediction errors is associated with less positive affect during task performance in the present experiment, the detected group difference might also reflect stronger emotional responsivity in individuals with depression. This would be in line with the idea of affected beliefs (Müller-Pinzler et al., 2022) and corresponds to the generally reduced positive affect in individuals with depression during the task. Accordingly, earlier studies found stronger reactions to negative emotional stimuli in individuals with depression together with reduced cognitive emotion regulation strategies like reappraisal (Joormann & Gotlib, 2010) and, therefore, include affective responses in models of depression (Smith et al., 2018).

The stronger neural and possibly affective reaction to more negative self-related feedback can also be discussed in the context of emotional schemas and their role in shaping emotional responses. Emotional schemas are formed by connecting highly arousing emotional events (Rouhani et al., 2023) to the resulting emotional reactions (Greenberg, 2010). They can be triggered by learned signs associated with the events, resulting in automatic, rapid, and exceptionally intensive emotional responses (Greenberg, 2010). Depression is more likely related to early maladaptive schemas rooted in a learning history of negative experiences, such as social devaluation in response to failure (Bishop et al., 2022; Rezaei et al., 2016). Therefore, individuals may have emotional schemas that are particularly sensitive to negative self-related feedback, which results in a stronger emotional response and, thus, an overall lower positive affect during task performance. This stronger reaction could be an expression of a stronger emphasis on these previously established negative internal models, making it more difficult to process external stimuli

that contradict them. To better understand task-specific negative affective reactions to self-related feedback, trial-by-trial fluctuations in happiness or other task-relevant self-conscious emotions like pride (Stolz et al., 2020) or embarrassment (Müller-Pinzler et al., 2015) should be measured in future studies.

Clinical implications of the neurocomputational findings: When patients are particularly sensitive to negative feedback, in addition to psychoeducation about emotions and emotional schemas, the finding of increased activation in the insula can be used to contextualize and validate the patient's emotional experiences. Subsequently, as mentioned above, skills to reduce emotional reactivity can be trained. Our findings of uncompromised processing of positive prediction errors in the VS or VTA/SN are in line with several other studies in depression, which show unaltered VS activity in reward learning tasks. Together with the above finding of similar learning in both groups, our results question the widespread hypothesis of altered VS reactivity and prediction error learning deficits in depression (Admon & Pizzagalli, 2015; Russo & Nestler, 2013; Whitton et al., 2015). The results suggest that the neural processing of positive self-related prediction errors as a basis for forming novel positive self-beliefs is unaltered, at least at this initial stage of processing in depression.

How can we interpret the reduced belief updating in the clinical sample following positive prediction errors while the neural tracking of positive prediction errors is unaffected?

A possible explanation is timing: the neural response is measured when feedback is given. The belief update may occur at a later stage, somewhere in between feedback, feedback processing, interpretation, and indication of the new expectation. As such, belief updates measured in behavior may represent an amalgam of several trials. In this case, cognitive appraisal mechanisms might have come into play. We also assume that stronger negative affective reactions in individuals with depression further promote cognitive distortions like cognitive immunization and, thus, influence self-ratings even in positive conditions.

The focus of the learning task used is on the formation of relatively novel beliefs (Krach et al., 2024). We selected estimation entities in which most people do not have strong prior beliefs regarding their ability. Although participants diagnosed with depression had lower overall self-esteem before the task, their prior beliefs about their estimation abilities were not significantly different. So, we can look at the underlying mechanisms of how people arrive at their beliefs and show that depressive symptomatology has an impact already at this stage. Although established, internal models can potentially influence the process of belief updating, the little prior experience that people typically have, e.g., in estimating the weight of animals, and thus the lower confidence in their self-beliefs prior to the task, is different from stable maladaptive beliefs that have been formed over time and continuously reinforced by symptom-maintaining behaviors. To experimentally address the revision of established beliefs to better map psychotherapy processes, future studies could use tasks that address more global negative self-beliefs with high confidence levels. Alternatively, the initial belief formation task could be followed by another session to challenge established beliefs again. Also, negative feedback during social interaction could be used to map the negative learning experience of many people with depression

more closely, which might activate maladaptive self-beliefs more strongly. This could further improve the translational nature of future work.

In conclusion, the present study sheds light on the neurocomputational mechanisms that contribute to maladaptive self-belief formation in depressive psychopathology. The results emphasize a heightened neural response to more negative self-related feedback in the insula and a potentially more emotional response that, together with less consideration of positive feedback, promotes more biased self-belief formation in depression (Figure 4.3). The study highlights the importance of accounting for within-group variance and symptom burden in a clinically relevant symptom range. Insights from the interaction between cognitive processes, affective experiences, and symptom burden towards forming maladaptive self-beliefs can be of great importance for future therapeutic intervention strategies.

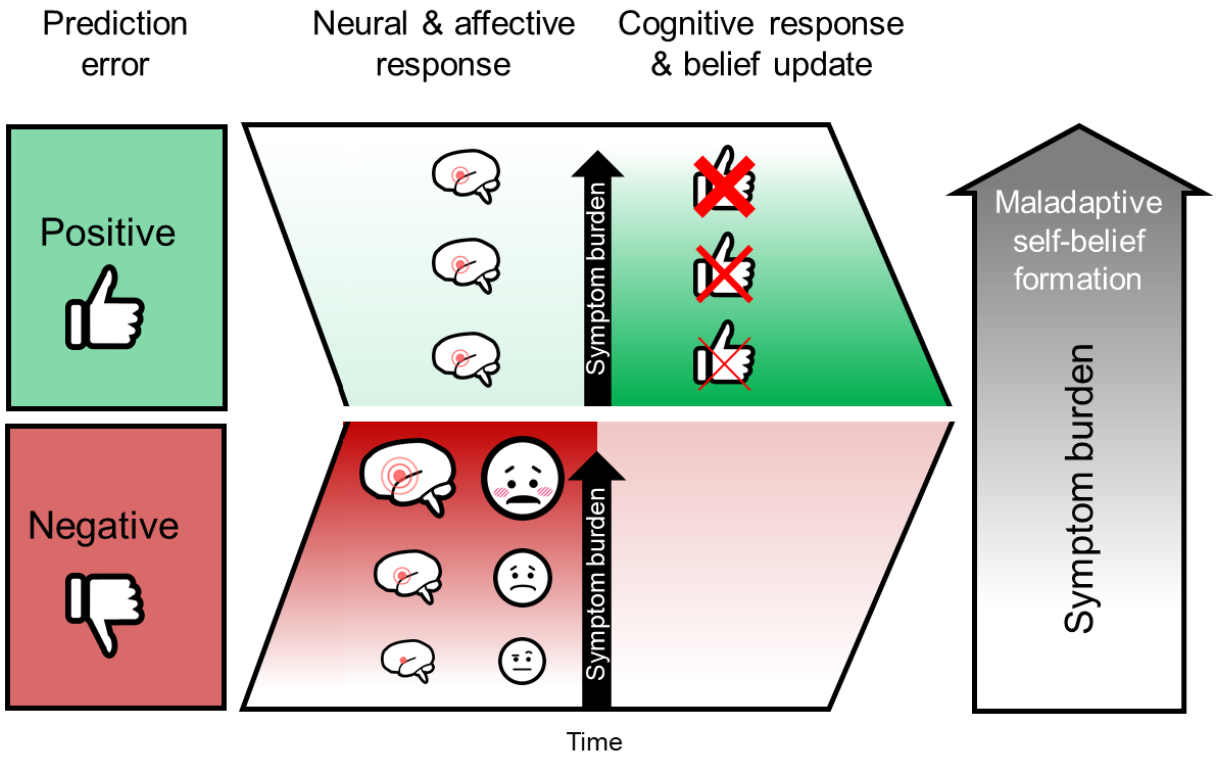


Figure 4.3. Schematic model of belief formation in relation to increasing depressive social anxiety symptom burden. Prediction error: Trial-by-trial positive and negative prediction errors were presented as social performance feedback in a belief updating task. Symptom burden: combined measure of Beck's depression inventory (Schmitt et al., 2003) as a general measure of depressive symptoms, Automatic Thought Questionnaire (Pössel et al., 2005) as a measure of cognitive symptoms of depression, as well as self-esteem (Marsh & O'neill, 1984), and social anxiety (Stangier et al., 1999) as transdiagnostic measures. Neural and affective response to feedback: Stronger activity within the insula with more negative prediction errors is associated with higher symptom burden. Stronger insula activity is also linked to less positive affect. Cognitive response and belief update: Fewer belief updates following positive prediction errors in the clinical sample might reflect the bias of cognitive immunization against unexpected positive feedback. Time: suggested interpretation of an initial neural and affective response during feedback presentation and possible downstream cognitive mechanisms with subsequent belief updating. All of this together might contribute to more biased beliefs with higher symptom burden.

4.5 Methods

Participants

The study was approved by the ethics committee of the University of Lübeck, was carried out in accordance with the ethical guidelines of the American Psychological Association, and all participants gave written informed consent. All participants were adults above the age of 18 and fluent in German. The goal in recruiting the clinical sample was to obtain a naturalistic sample with variance in depressive and social anxiety symptoms. Inclusion criteria were a diagnosis of depression (F32.1/2, F33.1/2); comorbid symptoms of social anxiety (F40.1, F60.6) were preferentially included but were not a necessary inclusion criterion. Exclusion criteria were acute suicidality, an acute psychotic state, schizophrenia, schizotypal disorder or delusional disorder, bipolar disorder, primary diagnosis of substance abuse or substance dependence, narcissistic, histrionic, or borderline personality disorder, and neurological disease. The clinical sample (n=35, 9 females, aged 20–55 years; M=34.23; SD=10.54) was recruited from the Department of Psychiatry and Psychotherapy at the university clinic Lübeck. Twenty-eight were outpatients, seven inpatients. Twenty-seven of the outpatients were in a day treatment program for depression (8 weeks, weekday mornings to afternoons) with multi-professional therapy. All patients received psychotherapy, and n = 28 received additional pharmacotherapy. Measurement day was, on average, 2.4 weeks (SD = 2.0) after admission to the clinic. When recruiting the control group, group-wise pairing of age and gender distribution was aimed at. Subjects of the control group (n=32, 10 females, aged 18–55 years; M=34.25; SD=10.23, no group differences in age or gender distribution p>.8, Supplementary Table 4.1) were recruited at the university campus of Lübeck or via public notices. The exclusion criterion was mental illness or neurological disease. In a telephone interview before the study, the subjects were screened for symptoms of depression, and in the second part, a general screening for other mental illnesses was conducted. Screenings were adapted from a structured interview (Wittchen et al., 1997). Two individuals had to be subsequently excluded despite screening due to clinically relevant BDI scores (> 35). We initially recruited 85 participants in total but had to exclude 18 participants. In the clinical sample, we excluded five due to technical problems, four due to premature termination of the study, and two participants who did not attentively complete the task until the end (e.g., ratings indicated that they stopped responding). In the control group, we excluded three participants who did not believe the cover story of the task and/ or did not attentively complete the task; one was excluded due to technical problems, one due to a premature termination of the study, and two due to high BDI scores as mentioned above.

Learning Of Own Performance task

The Learning Of Own Performance (LOOP) task allows participants to gradually learn about their own or another person's purported ability to estimate various properties through performance feedback after each trial. The task was previously introduced and validated in behavioral and fMRI studies on healthy participants (Czekalla et al., 2021; Müller-Pinzler et al., 2019, 2022). Two participants were invited to participate in the study

covered as an experiment on cognitive estimation; one performed the task in the scanner, and the other as a behavioral study (data not included). When a second participant was unavailable, a confederate was presented and allegedly placed in the adjacent room to participate in the behavioral study. The participants were informed that they would take turns with the other individual, either actively performing the task themselves (Self condition) or observing the other person's performance (Other condition). The task involved estimating different attributes in four estimation categories: house height, animal weight, vehicle distance, and food quantity. After each estimation trial, participants received manipulated performance feedback. In two estimation categories, the feedback was related to one's own performance and, in the other two, to the other person's estimation performance. One of the two estimation categories for each agent was associated with predominantly positive feedback (High Ability condition), while the other was linked to mostly negative feedback (Low Ability condition). Estimation categories were counterbalanced between Ability conditions and Agent conditions. This resulted in four feedback conditions (Agent [Self vs. Other] x Ability [High vs. Low]) with 20 trials each. The trials of all conditions were intermixed in a fixed order with a maximum of two consecutive trials of the same condition. Performance feedback presented after every estimation trial indicated the participant's or the other person's current estimation accuracy as percentiles compared to an alleged reference group of 350 former participants (e.g., "You are better than 94% of the reference participants"; see Figure 4.1). The feedback was defined by a sequence of predetermined prediction errors concerning the participants' current beliefs about their abilities. The current belief was calculated as the average of the last five performance expectation ratings per category, which started at 50% before participants rated their performance expectations. This procedure led to varying feedback sequences between participants while ensuring that the prediction errors remained mostly independent of the participants' performance expectations and that negative and positive prediction errors were relatively equally distributed across groups and conditions (MDD Self: mean positive PE = 14.2, *SD* = 1.5 (mean frequency = 20.4); mean negative PE = −12.8, *SD* = 1.8 (mean frequency = 19.1); MDD Other: mean positive PE = 13.6, *SD* = 1.3 (mean frequency = 19.0); mean negative PE = −14.0, *SD* = 1.8 (mean frequency = 20.3), CON Self: mean positive PE = 13.9, *SD* = 1.4 (mean frequency = 20.2); mean negative PE = −13.2, *SD* = 1.4 (mean frequency = 19.3); CON Other: mean positive PE = 13.8, *SD* = 1.7 (mean frequency = 18.7); mean negative PE = −13.5, *SD* = 1.2 (mean frequency = 20.7), no group differences in mean PE for all four PE conditions, all $p > 0.2$). At the beginning of each trial, a cue was presented indicating the estimation category (e.g., height) and the agent assigned for that trial (e.g., You). Participants were then asked to provide their expected performance for that specific trial on a scale with the same percentiles used for feedback. To enhance motivation and encourage honest responses, participants were informed, as part of the cover story, that accurate expected performance ratings would be rewarded with up to 6 cents per trial, that is, the closer their expected performance rating matched their actual feedback percentile, the more money they would receive. After each performance

expectation rating, the estimation question was presented for 10 seconds. During the estimation phase, continuous response scales under the pictures defined a range of plausible answers for each question. Participants indicated their responses by moving a pointer on the response scale using an MRI-compatible computer mouse. Subsequently, feedback was presented for 3 seconds (Figure 4.1A). Jittered inter-stimulus-intervals were presented following the cue (mean: 4*TR (0.992 s), range: 2–6*TR), estimation (mean: 4.5*TR, range: 2.5–6.5*TR), and feedback phase (mean: 6*TR, range: 4–8*TR) for the fMRI task with jitters distributed in a uniform distribution with steps of 0.5*TR. During the task, participants rated their current levels of embarrassment, pride, happiness, and stress/ arousal on a continuous scale ranging from not at all (coded as 0) to very strong (coded as 100). Emotion ratings were presented four times, each following a trial of one of the four experimental conditions (Self/Other, High/Low). The two happiness ratings following Self-trials were averaged to receive a measure of a general affective experience following self-related feedback. All stimuli were presented using MATLAB (Release 2015b, The MathWorks, Inc.) and the Psychophysics Toolbox (Brainard, D. H. [1997]. The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436). The fMRI task was completed in two separate approximately 25-minute sessions with a short break in between.

Before starting the experiment, demographic data and a self-esteem personality questionnaire (Self-Description Questionnaire-III, SDQ-III subscale scores; Marsh & O'neill, 1984) were retrieved. Following the task, participants completed questionnaires addressing the symptom burden, including Beck's depression inventory (BDI-V; Schmitt et al., 2003), Automatic Thought Questionnaire (ATQ; Pössel et al., 2005), and the Social Interaction Anxiety Scale (SIAS; Stangier et al., 1999).

Statistical analysis

Behavioral data analysis and modeling. First, we checked for group differences in prior beliefs (Supplementary Note 4.5, Supplementary Table 4.10) and other setting characteristics at the start (Supplementary Table 4.11). Then, we ran a model-free analysis on the participants' expected performance ratings for each trial. We employed a linear mixed model with a maximum likelihood estimation. The model included the factors Ability (High vs. Low) x Agent (Self vs. Other) x Group (MDD vs. CON) and Trial (20 trials) as continuous predictors. Intercept, Ability, Agent, Trial, and Group were modeled as fixed effects; the intercept was additionally modeled as a random effect (Supplementary Note 4.1, Supplementary Table 4.2).

After the model-free analyses, dynamic changes in performance expectation ratings were modeled using prediction error delta-rule update equations (adapted Rescorla–Wagner model, Rescorla & Wagner, 1972). The model space and the procedure of model fitting and selection have been implemented and further developed in our previous studies (Müller-Pinzler et al., 2019, 2022). For the learning models the following equation was used (EXP = Performance expectation rating, FB = feedback, PE = prediction error, α = learning rate):

$EXP_{t+1} = EXP_t + \alpha\ PE_t$; while $PE_t = FB_t - EXP_t$

The model space consisted of three main models, each with different assumptions regarding biased updating behavior when forming novel beliefs (for model space, see Supplementary Table 4.3). The simplest learning model (Unity Model) employed a single learning rate for all conditions for each participant, thus assuming no learning biases. The Valence Model included separate learning rates for positive (LR+) and negative prediction errors (LR-) across both ability conditions, suggesting that the valence (positive vs. negative) of prediction errors influences belief formation. The Ability Model incorporated distinct learning rates for each ability condition, indicating context-specific learning. In addition, learning rates were either estimated separately for Self vs. Other (Models 4, 5, 6, and 7) or across Agent conditions (Models 1, 2, and 3). The Valence Model with separate learning rates for Self vs. Other (Model 5), which was the winning model in our first studies (Müller-Pinzler et al., 2019), was further extended by adding a weighting factor. This factor reduced the learning rates when feedback values approached the extremes of the scale (percentiles close to 0% or 100%, Model 7), assuming that participants would perceive extreme feedback to be less likely than average feedback (Kube et al., 2022). Since many variables encountered in everyday life approximately follow a normal distribution where extreme values are less probable, we assigned the relative probability density of the normal distribution to each feedback percentile value. A weighting factor w was fitted for each individual, indicating how strongly the relative probability density reduced the learning rates for feedback further away from the mean. The Weighted Valence Model was the winning in our last study (Müller-Pinzler et al., 2022). The normal decay (ND) weighted by the weighting factor w was introduced in the learning models in the following way:

$EXP_{t+1} = EXP_t + \alpha\ PE_t\ (1 - w\ ND)$

The initial beliefs about the own and the other participant's performance (EXP_1) were estimated as free parameters separately for Self and Other in both Ability conditions, resulting in four additional model parameters. To compare whether the participants' performance expectation ratings can be better explained in terms of prediction error learning compared to assuming stable values within each ability condition, we included a simple Mean Model with a mean value for each task condition (Model 8).

Model fitting. For model fitting, we used the R package *'RStan'* (Stan Development Team [2022]. RStan: the R interface to Stan. R package version 2.21.7. https://mc-stan.org/). It employs Markov chain Monte Carlo (MCMC) sampling algorithms. Each participant's learning parameters for all models of the model space were individually fitted. A total of 2400 samples were drawn after 1000 burn-in samples (overall 3400 samples thinned with a factor of 3) in three MCMC chains. Posterior parameter distributions were sampled for each participant. We assessed whether MCMC chains converged by inspecting $\hat{R}$ values for all model parameters (Gelman & Rubin, 1992). Effective sample sizes n_{eff} of model parameters (effective number of independent draws from the posterior distribution) were typically greater than 1500 (for most parameters and subjects). To obtain a single

parameter value per participant for the group statistics, posterior distributions for all parameters for each participant were summarized by their mean.

Model selection. To select the model that best described the course of our performance expectation ratings, pointwise out-of-sample prediction accuracy was estimated by approximating leave-one-out cross-validation (LOO; Vehtari et al., 2017). LOOs were calculated for all fitted models separately for every participant using Pareto smoothed importance sampling (PSIS) with the log-likelihood from the posterior simulations of the parameter values (Vehtari et al., 2016). We calculated $\hat{k}$ values, the estimated shape parameters of the generalized Pareto distribution, a measure of the reliability of the PSIS-LOO (see Supplementary Table 4.4 for sum LOO Scores, $\hat{k}$ values, and difference scores of PSIS-LOO in contrast to the winning model). Only a few trials resulted in insufficient parameter values for $\hat{k}$ and thus potentially unreliable PSIS-LOO scores. Bayesian Model Selection on PSIS-LOO scores was performed on the group level for the whole sample and both sub-samples (Rigoux et al., 2014) using MATLAB (Release 2019b, The MathWorks, Inc.). It provides the protected exceedance probability for each model (*pxp*), indicating the likelihood of a given model explaining the data better than all other models in the comparison set. The Bayesian omnibus risk (*BOR*) quantifies the posterior probability that model frequencies for all models are equal. To assess whether the winning model captured the effects in the behavioral data, we repeated the model-free analysis (Supplementary Note 4.1) with the data predicted by the winning model; here, the Weighted Valence Model (Supplementary Note 4.3).

Capturing symptom burden. For a dimensional perspective on depression, we assessed the severity of depressive symptoms with the BDI-V as a general measure of depressive symptoms, as well as the ATQ as a more specific measure of the cognitive symptoms of depression. To address symptom burden in a broader context, we additionally assessed symptoms of social anxiety with the SIAS and self-esteem with SDQ-III subscale scores as a transdiagnostic measure since we could show a relationship with belief updating behavior and these two concepts in a previous study (Müller-Pinzler et al., 2019). To reduce dimensionality, we ran a principal component analysis on our questionnaire data using the *R* package *'psych'* (varimax rotation, component scores based upon the structure matrix [default], William Revelle (2023). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. *R* package version 2.3.3, https://CRAN.R-project.org/package=psych). We extracted one component as a combined measure of depressive and socially anxious psychopathology (Figure 4.1E).

Statistical analyses of learning parameters. To test whether the updating of self-beliefs was different between groups, learning rates of the winning models for positive (LR+) and negative prediction errors (LR-, factor Prediction error valence, PE_Val) and for Self and Other (factor Agent) were compared between the two groups in a PE_Val x Agent x Group linear mixed model. To associate learning biases with symptom burden, we computed a Valence Bias Score for Self and Other that captured a bias in belief updating as a

difference between the positive and negative learning rate, VBS=(LR[+] – LR[-])/ (LR[+] + LR[-]) (Müller-Pinzler et al., 2019, 2022; Niv et al., 2012), and calculated an Agent x Psychopathology score linear mixed-effects model. For a more detailed understanding of the relationship between biased belief updating and symptom burden, we additionally performed Spearman correlations between the Psychopathology score and the Valence Bias Score as well as the two separate self-related learning rates LR[+] and LR[-] within the two sub-samples. We compared the correlation of the Psychopathology score and the Valence Bias Score between the MDD and CON groups to see which group drives the effect. Since we had a particular interest in the relationship between the Psychopathology score and the self-related positive learning rate as a potential measure of cognitive immunization against positive feedback, we additionally tested the absolute correlation parameter of LR[+] and Psychopathology against the one with LR[-] within the clinical sample (*R* package *'cocor'* that implemented to correlation comparison approach by Pearson and Filon, 1898; Diedenhofen, B. & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE, 10*(4): e0121945). Statistical tests were performed two-sided. Statistical analyses on the behavioral data were performed using *R* (*R* Core Team [2022]. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/).

fMRI data acquisition. FMRI data were collected at the Center of Brain, Behavior, and Metabolism at the University of Lübeck, Germany, using a 3 T Siemens MAGNETOM Skyra scanner (Siemens, München, Germany) with 60 near-axial slices. In each of the two experimental sessions, 1516 functional volumes (min=1400, max=1724) were acquired on average using echo planar imaging (EPI, TR=0.992 s, TE=28 ms, flip angle=60°, voxel size= 3×3×3mm^3, simultaneous multi-slice factor 4). A high-resolution anatomical T1 image was obtained for normalization purposes (voxel size =1×1×1mm^3, 192 × 320 × 320 mm^3 field of view, TR=2.300 s, TE=2.94 ms, TI=900 ms; flip angle=9°; GRAPPA factor 2; acquisition time 6.55 min).

FMRI data analyses. FMRI data were preprocessed and analyzed with Statistical Parametric Mapping 12 (SPM12, Wellcome Trust Centre for Neuroimaging, University College London). Field maps were recorded to obtain voxel displacement maps (VDMs) to correct for geometric distortions. EPIs were slice-time corrected, motion-corrected, and unwarped using the corresponding VDMs, co-registered with the T1 image, and normalized using the forward deformation fields as obtained from the unified segmentation of the anatomical T1 image. The normalized volumes were resliced with a 2×2×2 mm^3 voxel size and smoothed using an 8mm, full-width-at-half-maximum isotropic Gaussian kernel. Functional images were high-pass filtered at 1/384 to remove low-frequency drifts. A two-level, mixed-effects procedure was implemented for the statistical analyses. On the first level, a fixed-effects GLM included four regressors for the cue conditions (Ability: High vs. Low × Agent: Self vs. Other), weighted with the performance expectation ratings per trial as parametric modulator, four regressors for the feedback conditions (Agent: Self vs.

Other × PE_Val: positive [+] vs. negative [-]), weighted with PE magnitude per trials (continuous effect of the unsigned PE values), one regressor for the performance expectation rating phase, two for the estimation period for Self and Other, and one for the emotion rating phase. Parametric modulators were not orthogonalized; thus, each only explained their specific variance. Regressors were modeled with the duration as presented during the experiment (cue phase: 2.5s, performance expectation rating: individual reaction times with *M*=4.0s, *SD*=2,4, estimation phase: 10s, feedback phase: 3s, emotion rating phase: *M*=26.2s, *SD*=10.1). Six additional regressors were included to correct for head movement, and one regressor was included with a constant term for each of the two sessions.

On the second level, we first compared the brain activity for self- vs. other-related positive and negative prediction errors in the feedback phase in a flexible factorial design (regressor for the four conditions Agent: Self x Other, PE_Val: PE[+] x PE[-], Supplementary Figure 4.4) and checked whether the brain activity in the positive self-related PE condition (categorical effect) correlates with $LR[+]_{Self}$ and for the negative PE condition with $LR[-]_{Self}$ within our predefined ROIs. For this, parameter estimates of all voxels within one ROI were extracted, averaged, and correlated.

Second, the tracking of positive and negative self-related prediction errors captured by the parametric modulator of PE magnitude (continuous effect) was compared between the MDD and CON group with a two-sample t-test within our ROIs (for the distribution of parameter estimates; see Supplementary Figure 4.8). To test whether the difference between the tracking of positive and negative prediction errors is stronger in the clinical compared to the control sample, we additionally checked the PE Valence x Group interaction in a flexible factorial design within our ROIs. To assess correlations with the Psychopathology score, the Valence Bias Score, and the happiness ratings, parameter estimates of all voxels within one ROI were extracted and averaged as mentioned above.

The selection of regions of interest was guided by our previous fMRI study with a healthy sample using the same paradigm (Müller-Pinzler et al., 2022). As the insula, with its suggested role as an integrative hub for motivated cognition and emotional behavior, has emerged as particularly relevant in this study, it was again defined as an ROI. We used unilateral insula ROIs with their dorsal, ventral, and posterior part as described in the three-cluster solution of Kelly and colleagues (Kelly et al., 2012). Other ROIs that were motivated by this study are in the amygdala (two unilateral ROIs from the AAL atlas definition in the WFU PickAtlas, Tzourio-Mazoyer et al., 2002) and an anatomically defined VTA/SN ROI, dopaminergic nuclei in the midbrain (probabilistic atlases of the midbrain; Adcock Lab; Ian C. Ballard, Vishnu P. Murty, R. McKell Carter, Jeffrey J. MacInnes, Scott A. Huettel and R. Alison Adcock, 2011; Murty et al., 2014). Since altered prediction error processing in the ventral striatum has been discussed a lot in the context of depression, a functional ROI within the ventral striatum (according to the SPM anatomy toolbox) capturing the processing of prediction error valence was additionally defined as ROI. It derived from a family-wise error $p<.05$ corrected baseline effect of the continuous signed self-related prediction errors of our previous study (Müller-Pinzler et al., 2022).

FMRI results were family-wise error (FWE) corrected at peak level for the whole brain or within the ROIs. Coordinates are reported in the MNI space. Anatomical labels of all resulting clusters were derived from the SPM Anatomy toolbox, version 3.0 (Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. [2005]. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage, 25(4), 1325-1335).

4.6 References

Adams, G. C., Balbuena, L., Meng, X., & Asmundson, G. J. G. (2016). When social anxiety and depression go together: A population study of comorbidity and associated consequences. *Journal of Affective Disorders*, *206*, 48–54.

Admon, R., & Pizzagalli, D. A. (2015). Dysfunctional reward processing in depression. *Current Opinion in Psychology*, *4*, 114–118.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (5th ed., text rev.)*. American Psychiatric Association Publishing.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bandura, A., Pastorelli, C., Barbaranelli, C., & Caprara, G. V. (1999). Self-efficacy pathways to childhood depression. *Journal of Personality and Social Psychology*, *76*(2), 258–269.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(11), 1833.

Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, *9*(4), 324–333.

Beck, A. T. (1964). Thinking and depression: II. Theory and therapy. *Archives of General Psychiatry*, *10*(6), 561–571.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. International Universities Press.

Beck, Aaron T. (1979). *Cognitive Therapy of Depression*. Guilford Press.

Bennett, D., Radulescu, A., Zorowitz, S., Felso, V., & Niv, Y. (2023). Affect-congruent attention modulates generalized reward expectations. *PLoS Computational Biology*, *19*(12), e1011707.

Bishop, A., Younan, R., Low, J., & Pilkington, P. D. (2022). Early maladaptive schemas and depression in adulthood: A systematic review and meta-analysis. *Clinical Psychology & Psychotherapy*, *29*(1), 111–130.

Brolsma, S. C. A., Vrijsen, J. N., Vassena, E., Rostami Kandroodi, M., Bergman, M. A., van Eijndhoven, P. F., Collard, R. M., den Ouden, H. E. M., Schene, A. H., & Cools, R. (2022). Challenging the negative learning bias hypothesis of depression: reversal learning in a naturalistic psychiatric sample. *Psychological Medicine*, *52*(2), 303–313.

Bromberg-Martin, E. S., & Sharot, T. (2020). The value of beliefs. *Neuron*, *106*(4), 561–565.

Brotzeller, F., & Gollwitzer, M. (2024). Exploring Asymmetries in Self-Concept Change After Discrepant Feedback. *Personality & Social Psychology Bulletin*, 1461672241232738.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience and Biobehavioral Reviews*, *55*, 247–267.

Clark, J. E., Watson, S., & Friston, K. J. (2018). What is mood? A computational perspective. *Psychological Medicine*, *48*(14), 2277–2284.

Czekalla, N., Stierand, J., Stolz, D. S., Mayer, A. V., Voges, J. F., Rademacher, L., Paulus, F. M., Krach, S., & Müller-Pinzler, L. (2021). Self-beneficial belief updating as a coping mechanism for stress-induced negative affect. *Scientific Reports*, *11*(1), 1–13.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196.

Diederen, K. M. J., Spencer, T., Vestergaard, M. D., Fletcher, P. C., & Schultz, W. (2016). Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron*, *90*(5), 1127–1138.

Dobson, K. S., & Dozois, D. J. A. (2021). *Handbook of Cognitive-Behavioral Therapies, Fourth Edition*. Guilford Publications.
Donaldson, C., Lam, D., & Mathews, A. (2007). Rumination and attention in major depression. *Behaviour Research and Therapy*, *45*(11), 2664–2678.
Eil, D., & Rao, J. M. (2010). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138.
Eizenman, M., Yu, L. H., Grupp, L., Eizenman, E., Ellenbogen, M., Gemar, M., & Levitan, R. D. (2003). A naturalistic visual scanning approach to assess selective attention in major depressive disorder. *Psychiatry Research*, *118*(2), 117–128.
Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*, 6149.
Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24.
Engelmann, J. B., Berns, G. S., & Dunlop, B. W. (2017). Hyper-responsivity to losses in the anterior insula during economic choice scales with depression severity. *Psychological Medicine*, *47*(16), 2879–2891.
Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545.
Everaert, J., Bronstein, M. V., Cannon, T. D., & Joormann, J. (2018). Looking through tinted glasses: depression and social anxiety are related to both interpretation biases and inflexible negative interpretations. *Clinical Psychological Science*, *6*(4), 517–528.
Everaert, J., Bronstein, M. V., Castro, A. A., Cannon, T. D., & Joormann, J. (2020). When negative interpretations persist, positive emotions don't! Inflexible negative interpretations encourage depression and social anxiety by dampening positive emotions. *Behaviour Research and Therapy*, *124*, 103510.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836.
Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Frontiers in Human Neuroscience*, *8*, 639.
Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *37*(7), 1297–1310.
Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–511.
Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J., & Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain: A Journal of Neurology*, *134*(6), 1751–1764.
Greenberg, L. S. (2010). Emotion-Focused Therapy: A Clinical Synthesis. *FOCUS*, *8*(1), 32–42.
Gross, J. J., & Muñoz, R. F. (1995). Emotion regulation and mental health. *Clinical Psychology: Science and Practice*, *2*(2), 151–164.
Harlé, K. M., Chang, L. J., van 't Wout, M., & Sanfey, A. G. (2012). The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. *NeuroImage*, *61*(1), 32–40.
Hayes, S., Strosahl, K., & Wilson, K. (1999). *Acceptance and commitment therapy* (Vol. 6). New York: Guilford press.
Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2014). Chapter 24 - A Cognitive-Behavioral Model of Social Anxiety Disorder. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social Anxiety (Third Edition)* (pp. 705–728). Academic Press.
Heitmann, C. Y., Peterburs, J., Mothes-Lasch, M., Hallfarth, M. C., Böhme, S., Miltner, W. H. R., & Straube, T. (2014). Neural correlates of anticipation and processing of performance feedback in social anxiety. *Human Brain Mapping*, *35*(12), 6023–6031.
Hollon, S. D., & Kendall, P. C. (1980). Cognitive self-statements in depression: Development of an automatic thoughts questionnaire. *Cognitive Therapy and Research*, *4*(4), 383–395.
Hoven, M., Luigjes, J., Denys, D., Rouault, M., & van Holst, R. J. (2023). How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach. *Nature Mental Health*, *1*(5), 337–345.

Hyman, S. E. (2021). Psychiatric disorders: grounded in human biology but not natural kinds. *Perspectives in Biology and Medicine*, *64*(1), 6–28.

Ian C. Ballard, Vishnu P. Murty, R. McKell Carter, Jeffrey J. MacInnes, Scott A. Huettel and R. Alison Adcock. (2011). Dorsolateral prefrontal cortex drives mesolimbic dopaminergic regions to initiate motivated behavior. *J. Neurosci.*, *31*(28), 10340–10346.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, *167*(7), 748–751.

Joormann, J., & Gotlib, I. H. (2010). Emotion regulation in depression: relation to cognitive inhibition. *Cognition & Emotion*, *24*(2), 281–298.

Joormann, J., & Quinn, M. E. (2014). Cognitive processes and emotion regulation in depression. *Depression and Anxiety*, *31*(4), 308–315.

Karnick, A. T., Bauer, B. W., & Capron, D. W. (2024). Negative mood and optimism bias: An experimental investigation of sadness and belief updating. *Journal of Behavior Therapy and Experimental Psychiatry*, 101910.

Kelly, C., Toro, R., Di Martino, A., Cox, C. L., Bellec, P., Castellanos, F. X., & Milham, M. P. (2012). A convergent functional architecture of the insula emerges across imaging modalities. *NeuroImage*, *61*(4), 1129–1142.

Kessler, R. C., Stang, P., Wittchen, H. U., Stein, M., & Walters, E. E. (1999). Lifetime co-morbidities between social phobia and mood disorders in the US National Comorbidity Survey. *Psychological Medicine*, *29*(3), 555–567.

Kessler, Ronald C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 617–627.

Klayman, J. (1995). Varieties of Confirmation Bias. *Psychology of Learning and Motivation*, *32*, 385–418.

Koban, L., Andrews-Hanna, J. R., Ives, L., Wager, T. D., & Arch, J. J. (2023). Brain mediators of biased social learning of self-perception in social anxiety disorder. *Translational Psychiatry*, *13*(1), 292.

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neuroscience and Biobehavioral Reviews*, *46 Pt 1*, 71–84.

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844.

Korn, Christoph W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592.

Koster, E. H. W., De Raedt, R., Goeleven, E., Franck, E., & Crombez, G. (2005). Mood-congruent attentional bias in dysphoria: maintained attention to and impaired disengagement from negative information. *Emotion* , *5*(4), 446–455.

Krach, S., Müller-Pinzler, L., Czekalla, N., Schröder, A., Lübber, F., Rademacher, L., Stolz, D. S., Paulus, F. M., Wilhelm, I., & Mayer, A. V. (2024). Examining self-belief formation through KI beliefs. In *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/2y5tv

Kube, T. (2023). Biased belief updating in depression. *Clinical Psychology Review*, *103*, 102298.

Kube, T., Kirchner, L., Gärtner, T., & Glombiewski, J. A. (2023). How negative mood hinders belief updating in depression: results from two experimental studies. *Psychological Medicine*, *53*(4), 1288–1301.

Kube, T., Kirchner, L., Lemmer, G., & Glombiewski, J. A. (2022). How the Discrepancy Between Prior Expectations and New Information Influences Expectation Updating in Depression—The Greater, the Better? *Clinical Psychological Science*, *10*(3), 430–449.

Kube, T., Rief, W., Gollwitzer, M., Gärtner, T., & Glombiewski, J. A. (2019). Why dysfunctional expectations in depression persist - Results from two experimental studies investigating cognitive immunization. *Psychological Medicine*, *49*(9), 1532–1544.

Kube, T., Schwarting, R., Rozenkrantz, L., Glombiewski, J. A., & Rief, W. (2020). Distorted cognitive processes in major depression: a predictive processing perspective. *Biological Psychiatry*, *87*(5), 388–398.

Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., & Steele, J. D. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain: A Journal of Neurology*, *131*(Pt 8), 2084–2093.

Kumar, Poornima, Goer, F., Murray, L., Dillon, D. G., Beltzer, M. L., Cohen, A. L., Brooks, N. H., & Pizzagalli, D. A. (2018). Impaired reward prediction error encoding and striatal-midbrain

connectivity in depression. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *43*(7), 1581–1588.
Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). *NeuroImage The role of the neural reward circuitry in self-referential optimistic belief updates*. *133*, 151–162.
Lewinsohn, P. M. (1974). A behavioral approach to depression. *Essential Papers on Depression*, 150–172.
Lockwood, P. L., & Klein-Flügge, M. C. (2020). Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, *16*(8), 761–771.
Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology. General*, *105*(1), 3–46.
Markus, H. R., & Wurf, E. (1987). The dynamic self-concept: a social psychological perspective. *Annual Review of Psychology*, *38*(1), 299–337.
Marsh, H. W., & O'neill, R. (1984). Self Description Questionnaire iii: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, *21*(2), 153–174.
McCarthy, P. A., & Morina, N. (2020). Exploring the association of social comparison with depression and anxiety: A systematic review and meta-analysis. *Clinical Psychology & Psychotherapy*, *27*(5), 640–671.
McHugh, S. B., Barkus, C., Huber, A., Capitão, L., Lima, J., Lowry, J. P., & Bannerman, D. M. (2014). Aversive prediction error signals in the amygdala. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(27), 9024–9033.
Morrison, K. H., Bradley, R., & Westen, D. (2003). The external validity of controlled clinical trials of psychotherapy for depression and anxiety: a naturalistic study. *Psychology and Psychotherapy*, *76*(Pt 2), 109–132.
Mulej Bratec, S., Xie, X., Schmid, G., Doll, A., Schilbach, L., Zimmer, C., Wohlschläger, A., Riedl, V., & Sorg, C. (2015). Cognitive emotion regulation enhances aversive prediction error activity while reducing emotional responses. *NeuroImage*, *123*, 138–148.
Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., Einhäuser, W., Paulus, F. M., & Krach, S. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, *119*(0), 252–261.
Müller-Pinzler, Laura, Czekalla, N., Mayer, A. V., Schröder, A., Stolz, D. S., Paulus, F. M., & Krach, S. (2022). Neurocomputational mechanisms of affected beliefs. *Communications Biology*, *5*(1), 1241.
Müller-Pinzler, Laura, Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., Paulus, F. M., & Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, *9*(1), 14416.
Murray, E. A. (2007). The amygdala, reward and emotion. *Trends in Cognitive Sciences*, *11*(11), 489–497.
Murty, V. P., Shermohammed, M., Smith, D. V., Carter, R. M., Huettel, S. A., & Adcock, R. A. (2014). Resting state networks distinguish human ventral tegmental area from substantia nigra. *NeuroImage*, *100*, 580–589.
Mutschler, I., Ball, T., Wankerl, J., & Strigo, I. A. (2012). Pain and emotion in the insular cortex: evidence for functional reorganization in major depression. *Neuroscience Letters*, *520*(2), 204–209.
Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562.
Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, *26*(7), 607–621.
Phelps, E. A. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53.
Pössel, P., Seemann, S., & Hautzinger, M. (2005). Evaluation eines deutschsprachigen Instrumentes zur Erfassung positiver und negativer automatischer Gedanken [Evaluation of a German-language instrument for recording positive and negative automatic thoughts]. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *34*(1), 27–34.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rezaei, M., Ghazanfari, F., & Rezaee, F. (2016). The role of childhood trauma, early maladaptive schemas, emotional schemas and experimental avoidance on depression: A structural equation modeling. *Psychiatry Research*, *246*, 407–414.

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., & Thorwart, A. (2015). Expectancies as core features of mental disorders. *Current Opinion in Psychiatry*, *28*(5), 378–385.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*, 971–985.

Robinson, O. J., Cools, R., Carlisi, C. O., Sahakian, B. J., & Drevets, W. C. (2012). Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. *The American Journal of Psychiatry*, *169*(2), 152–159.

Roefs, A., Fried, E. I., Kindt, M., Martijn, C., Elzinga, B., Evers, A. W. M., Wiers, R. W., Borsboom, D., & Jansen, A. (2022). A new science of mental disorders: Using personalised, transdiagnostic, dynamical systems to understand, model, diagnose and treat psychopathology. *Behaviour Research and Therapy*, *153*, 104096.

Rothkirch, M., Tonn, J., Köhler, S., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain: A Journal of Neurology*, *140*(4), 1147–1157.

Rouhani, N., & Niv, Y. (2019). Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, *236*(8), 2425–2435.

Rouhani, N., Niv, Y., Frank, M. J., & Schwabe, L. (2023). Multiple routes to enhanced memory for emotionally relevant events. *Trends in Cognitive Sciences*, *27*(9), 867–882.

Russo, S. J., & Nestler, E. J. (2013). The brain reward circuitry in mood disorders. *Nature Reviews. Neuroscience*, *14*(9), 609–625.

Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., Lam, J., Skandali, N., Siegel, J. Z., Ousdal, O. T., Prabhu, G., Dayan, P., Fonagy, P., & Dolan, R. J. (2017). Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry* , *74*(8), 790–797.

Safra, L., Chevallier, C., & Palminteri, S. (2019). Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Computational Biology*, *15*(7), e1007224.

Samoilov, A., & Goldfried, M. R. (2000). Role of emotion in cognitive-behavior therapy. *Clinical Psychology: A Publication of the Division of Clinical Psychology of the American Psychological Association*, *7*(4), 373–385.

Schmitt, M., Beckmann, M., Dusi, D., Maes, J., Schiller, A., & Schonauer, K. (2003). Messgüte des vereinfachten Beck-Depressions-Inventars (BDI-V) [Measurement quality of the simplified Beck Depression Inventory (BDI-V)]. *Diagnostica*, *49*(4), 147–156.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27.

Sears, C. R., Newman, K. R., Ference, J. D., & Thomas, C. L. (2011). Attention to emotional images in previously depressed individuals: an eye-tracking study. *Cognitive Therapy and Research*, *35*(6), 517–528.

Segal, Z., Williams, M., & Teasdale, J. (2018). *Mindfulness-Based Cognitive Therapy for Depression, Second Edition*. Guilford Publications.

Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How local and global metacognition shape mental health. *Biological Psychiatry*, *90*(7), 436–446.

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, *8*(9), 1234–1240.

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479.

Sliz, D., & Hayley, S. (2012). Major depressive disorder and alterations in insular cortical activity: a review of current functional magnetic imaging research. *Frontiers in Human Neuroscience*, *6*, 323.

Smith, R., Alkozei, A., Killgore, W. D. S., & Lane, R. D. (2018). Nested positive feedback loops in the maintenance of major depression: An integration and extension of previous models. *Brain, Behavior, and Immunity*, *67*, 374–397.

Stangier, U., Heidenreich, T., Berardi, A., Golbs, U., & Hoyer, J. (1999). Die Erfassung sozialer Phobie durch die Social Interaction Anxiety Scale (SIAS) und die Social Phobia Scale (SPS) [The assessment of social phobia using the Social interaction anxiety scale (SIAS) and the social phobia scale (SPS)]. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *28*(1), 28–36.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, *10*, 550.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643.

Stolz, D. S., Müller-Pinzler, L., Krach, S., & Paulus, F. M. (2020). Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nature Communications 2020 11:1*, *11*(1), 1–13.

Swallow, S. R., & Kuiper, N. A. (1988). Social comparison and negative self-evaluations: An application to depression. *Clinical Psychology Review*, *8*(1), 55–76.

Touroutoglou, A., Hollenbeck, M., Dickerson, B. C., & Feldman Barrett, L. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *NeuroImage*, *60*(4), 1947–1958.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289.

Ubl, B., Kuehner, C., Kirsch, P., Ruttorf, M., Diener, C., & Flor, H. (2014). Altered neural reward and loss processing and prediction error signalling in depression. *Social Cognitive and Affective Neuroscience*, *10*(8), 1102–1112.

Vandendriessche, H., & Palminteri, S. (2023). Neurocognitive biases from the lab to real life. *Communications Biology*, *6*(1), 158.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *J. Mach. Learn. Res.*, *17*, 3581–3618.

Villano, W. J., & Heller, A. S. (2024). Depression is associated with blunted affective responses to naturalistic reward prediction errors. *Psychological Medicine*, 1–9.

Villano, W. J., Kraus, N. I., Reneau, T. R., Jaso, B. A., Otto, A. R., & Heller, A. S. (2023). Individual differences in naturalistic learning link negative emotionality to the development of anxiety. *Science Advances*, *9*(1), eadd2976.

Wächter, T., Lungu, O. V., Liu, T., Willingham, D. T., & Ashe, J. (2009). Differential effect of reward and punishment on procedural learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(2), 436–443.

Waltz, J. A., Schweitzer, J. B., Gold, J. M., Kurup, P. K., Ross, T. J., Salmeron, B. J., Rose, E. J., McClure, S. M., & Stein, E. A. (2009). Patients with schizophrenia have a reduced neural response to both unpredictable and predictable primary reinforcers. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *34*(6), 1567–1577.

Weary, G., Elbin, S., & Hill, M. G. (1987). Attributional and social comparison processes in depression. *Journal of Personality and Social Psychology*, *52*(3), 605–610.

Wells, A. (2011). *Metacognitive Therapy for Anxiety and Depression*. Guilford Press.

Whitton, A. E., Treadway, M. T., & Pizzagalli, D. A. (2015). Reward processing dysfunction in major depression, bipolar disorder and schizophrenia. *Current Opinion in Psychiatry*, *28*(1), 7–12.

Will, G.-J., Moutoussis, M., Womack, P. M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., NSPN Consortium, Rutledge, R. B., & Dolan, R. J. (2020). Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Translational Psychiatry*, *10*(1), 96.

Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *ELife*, *6*, 1–21.

Wittchen, H.-U., Zaudig, M., & Fydrich, T. (1997). *SKID. Strukturiertes Klinisches Interview für DSM-IV. Achse I und II. Handanweisung*. Hogrefe.

Zamfir, E., & Dayan, P. (2022). Interactions between attributions and beliefs at trial-by-trial level: Evidence from a novel computer game task. *PLoS Computational Biology*, 18(9), e1009920.

4.7 Supplementary Information

Supplementary Notes

Supplementary Note 4.1: *Model-free behavioral analyses.*

We performed a model-free analysis on our performance expectation ratings over time to capture the basic effects of belief updating in our observed data and compare it between the two groups. The Trial x Ability condition x Agent condition x Group linear mixed model showed a significant main effect of Ability condition ($t(5279) = 4.4$, $p < .001$) and interaction of Trial x Ability condition ($t(55279) = 10.3$, $p < .001$), indicating that participants adapted their performance expectation ratings according to the presented feedback in each Ability condition (see Figure 4.1C). Moreover, there was a significant main effect of Agent ($t(5279) = -2.9$, $p = .004$) and a significant interaction of Trial x Agent ($t(5279) = -2.6$, $p = .010$), indicating that participants evaluated their performance increasingly more negatively over time than the other's performance. The main effect of Group and all interactions including Group were not significant (for a full results table, see Supplementary Table 4.2).

Supplementary Note 4.2: *Model comparison.*

Model 8, including separate learning rates for positive and negative prediction errors for Self vs. Other, received the highest sum PSIS-LOO score (approximate leave-one-out cross-validation [LOO] using Pareto smoothed importance sampling [PSIS], Vehtari et al., 2016) out of all models (for the structure of the model space see Supplementary Table 4.3, for all PSIS-LOO scores, see Supplementary Table 4.4). In addition, Bayesian model selection (Rigoux et al., 2014) in the whole sample and the clinical and control sub-sample resulted in a protected exceedance probability of $pxp > .999$ for this model and a Bayesian Omnibus Risk of $BOR < .001$. The expected model frequency was 55.07 (MDD: 27.51, CON: 27.48). Thus, the Weighted Valence Model was selected for all further analyses of learning parameters, allowing for a comparison of valence-specific learning rates (for parameter correlations, see Supplementary Figure 4.10).

Supplementary Note 4.3: *Posterior predictive checks: Behavioral analyses on the predicted data.*

We repeated the analyses (Supplementary Note 4.1) with the predicted data from the winning model to see whether our winning model captured the core effects in our model-free analysis. We could reproduce all effects from the model-free data with the predicted data (main effect Ability condition: $t(5279) = 4.04$, $p < .001$, main effect Agent condition: $t(5279) = -4.09$, $p < .001$, interaction Trial x Ability condition: $t(5279) = 14.2$, $p < .001$, interaction Trial x Agent condition $t(5279) = -2.51$, $p = 0.012$). This confirms that the winning model recapitulates the main effects in our data.

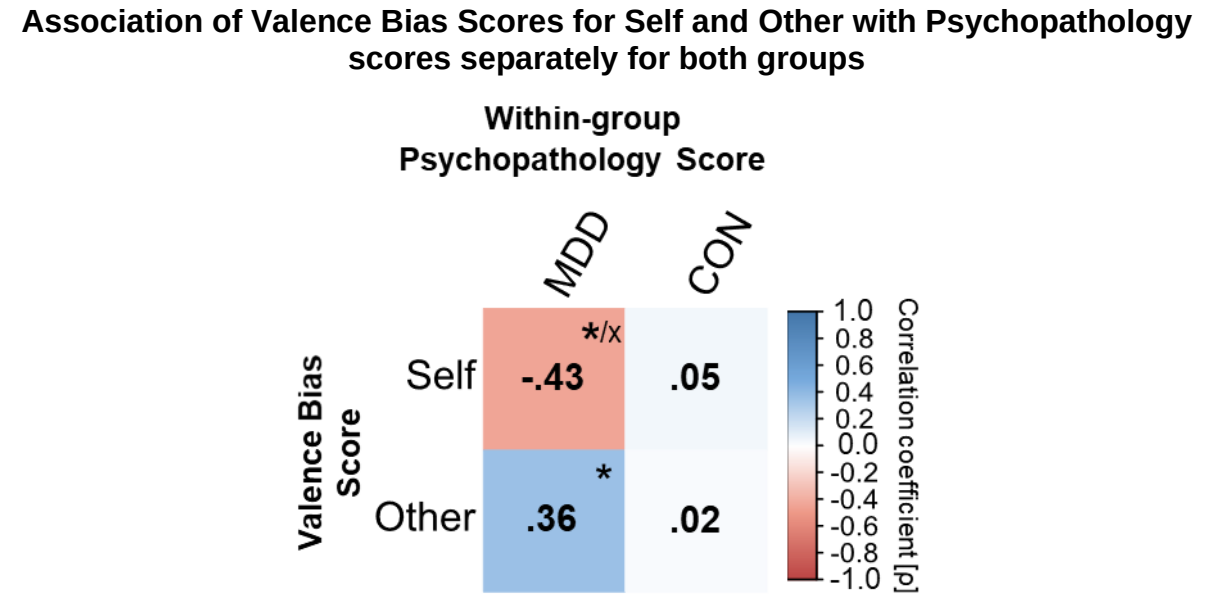
Supplementary Note 4.4: Stronger correlation between biased belief updating and symptom severity within the clinical sample compared to control.

Following the significant Agent x Psychopathology interaction (see Results section), a closer look at the distribution of the Psychopathology score and biased learning within groups suggests that the effect of increasingly biased updating behavior with increasing symptom severity is mainly driven by the clinical sample. Here, the correlation coefficients for the self-related Valence Bias Score indicate a significantly more negative relationship with levels of psychopathology in the clinical sample compared to the control group (ρ_{MDD} = -0.43, ρ_{CON} = 0.05, z = -1.99, p = 0.047). This difference does not reach significance for the other-related Valence Bias Score (ρ_{MDD} = 0.36, ρ_{CON} = 0.02, z = 1.4, p = 0.161). This demonstrates the importance of including individuals with clinically relevant symptom severity in the samples when studying psychopathological syndromes.

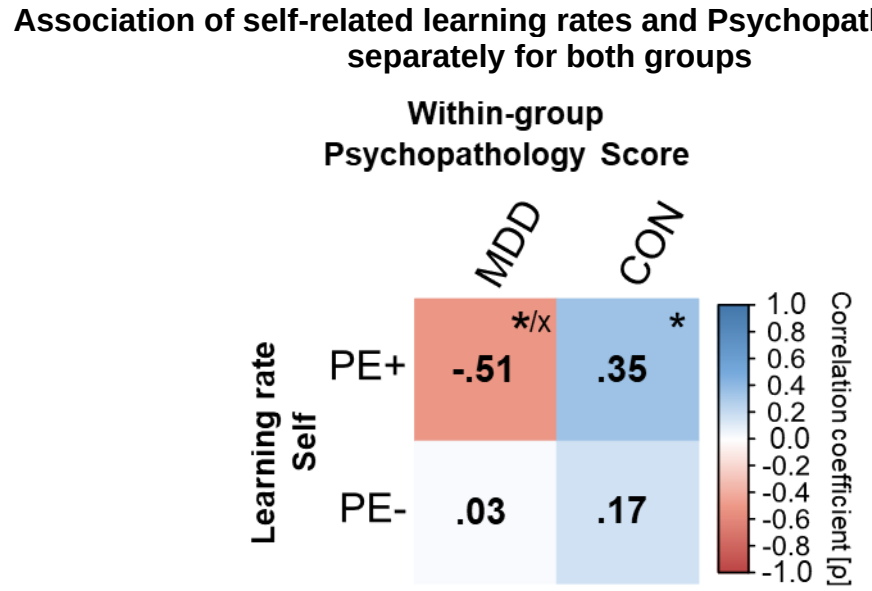
Supplementary Note 4.5: Group differences in global but not in specific prior beliefs.

Before comparing the process of belief formation between the two groups, we checked if there were already group differences in participants' prior beliefs regarding their estimation abilities. Individuals with depression showed a lower evaluation of their abilities in general (*SDQ-III*, t(45.06) = 8.23, p <. 001) as well as their general estimation ability before the task (t(62.15) = 2.1, p = 0.04). However, when asking more specifically for the performance in a certain estimation category in the upcoming first trial, there were no group differences, neither for self (t(55.82) = 1.83, p = .072) nor for other-related prior beliefs (t(59.35) = 1.39, p = .17). This allows for a comparison of the learning process between groups independently of the prior beliefs (for a complete table of prior beliefs with group comparisons see Supplementary Table 4.10).

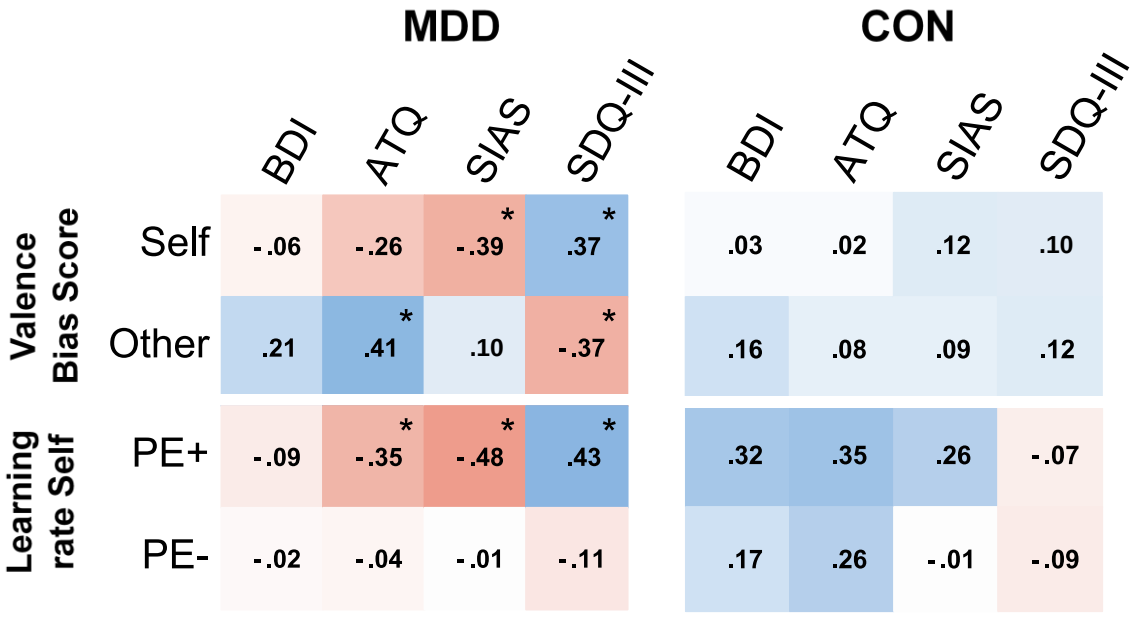
Supplementary Figures



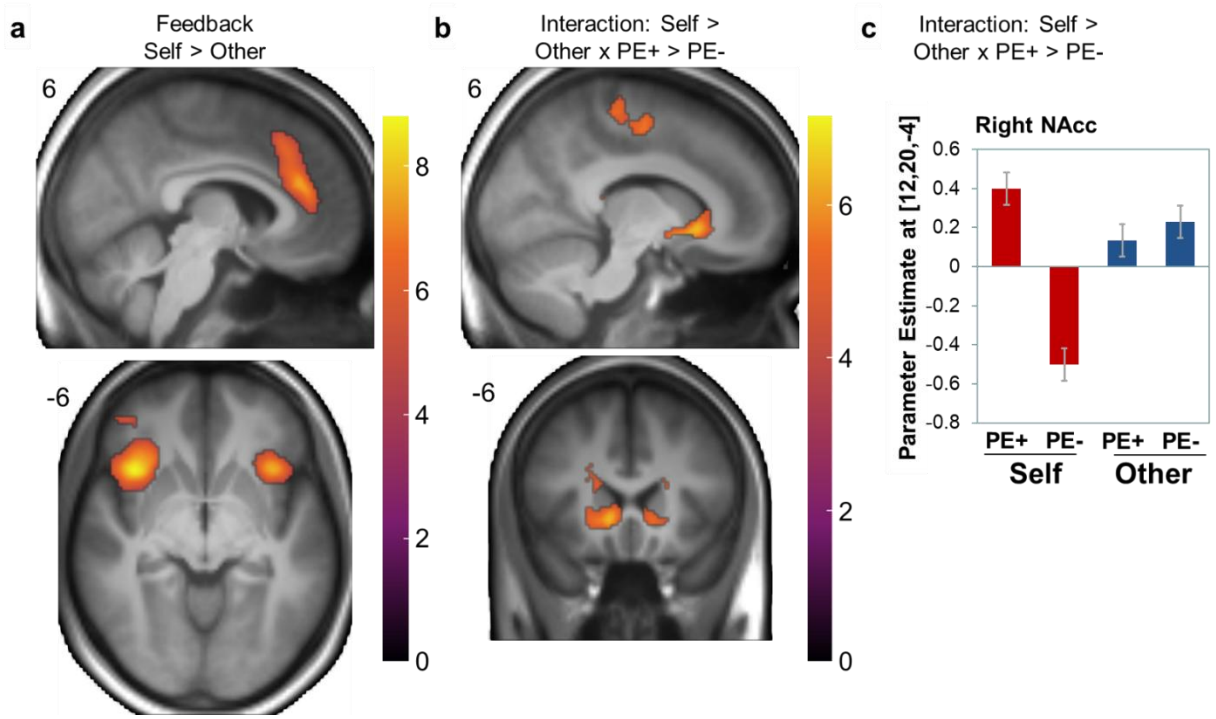
Supplementary Figure 4.1. Spearman correlation between Valence Bias Score (LR+ - LR-)/ (LR+ + LR-) for Self and Other with Psychopathology score for MDD and CON group. MDD = Major depressive disorder, CON = Control, * p < .05, x FDR corrected.



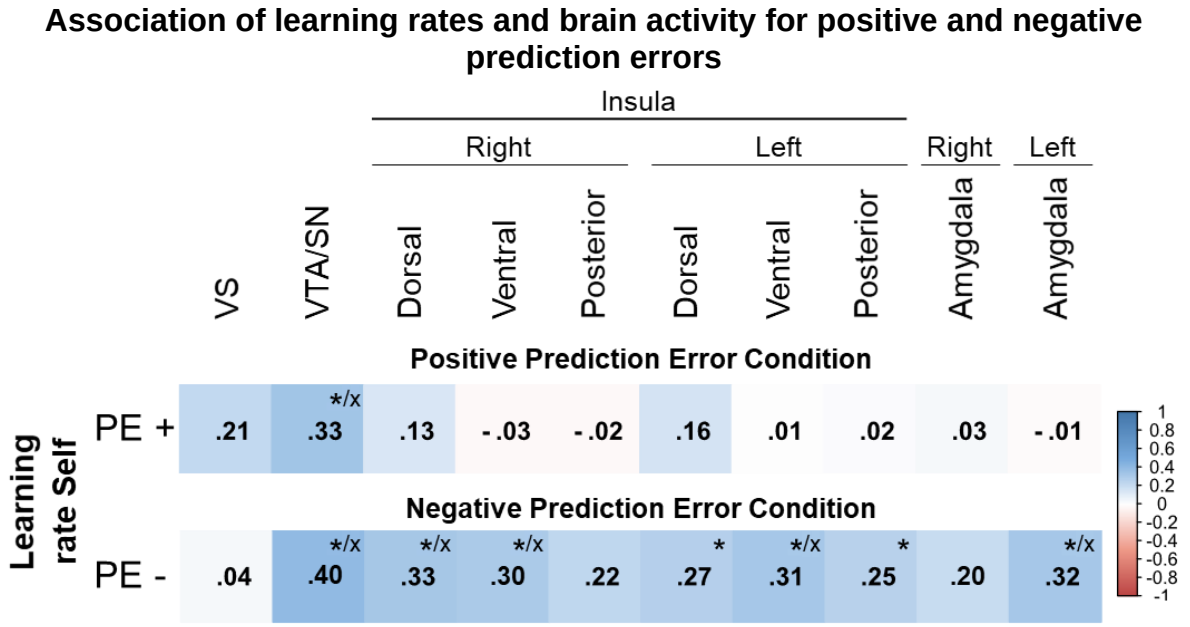
Supplementary Figure 4.2. Spearman correlation between self-related learning rates for positive (PE+) and negative prediction errors (PE-) with Psychopathology score for MDD and CON group. MDD = Major depressive disorder, CON = Control, * $p < .05$, x FDR corrected.



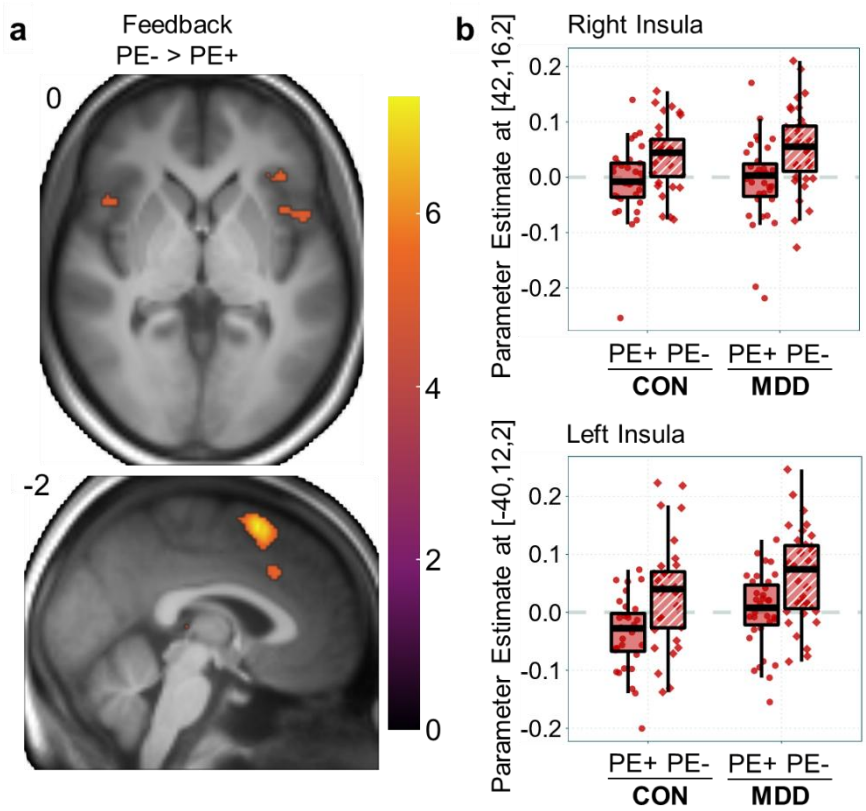
Supplementary Figure 4.3. Spearman correlation between the Valence Bias Score (LR+ - LR-)/ (LR+ + LR-) for Self and Other as well as the self-related learning rates for positive (PE+) and negative prediction errors (PE-) with the separate elements of the Psychopathology score: Beck's depression inventory (BDI-V, Schmitt et al., 2003), Automatic Thought Questionnaire (ATQ, Pössel et al., 2005), Social Interaction Anxiety scale (SIAS, Stangier et al., 1999), Self-Description Questionnaire-III (SDQ-III subscale scores, Marsh & O'neill, 1984). MDD = Major depressive disorder, CON = Control, * $p < .05$.



Supplementary Figure 4.4. Neural activations associated with prediction error processing (*categorical PE effect)* in the feedback phase. Replication of a previous study (Müller-Pinzler et al., 2022). a) Self-related feedback vs. other-related feedback was associated with increased activation of the bilateral insula cortex, frontal orbital cortex, cingulate gyrus, supramarginal gyrus ($p < .05$, FWE corrected at peak level for the whole brain). b) The interaction of Agent and Prediction Error Valence ([Self PE+ > Self PE-] **>** [Other PE+ > Other PE-]) resulted in activation of the bilateral VS, precentral/ postcentral gyrus, left hippocampus ($p < .05$, FWE corrected at peak level for the whole brain). Anatomical labels were derived from the SPM Anatomy Toolbox Version 3.0. c) Means and standard errors of parameter estimates corresponding to the BOLD response to positive and negative PEs in the right VS: More activity in VS for positive relative to negative PEs only when seeing self-related feedback, but not when observing others.

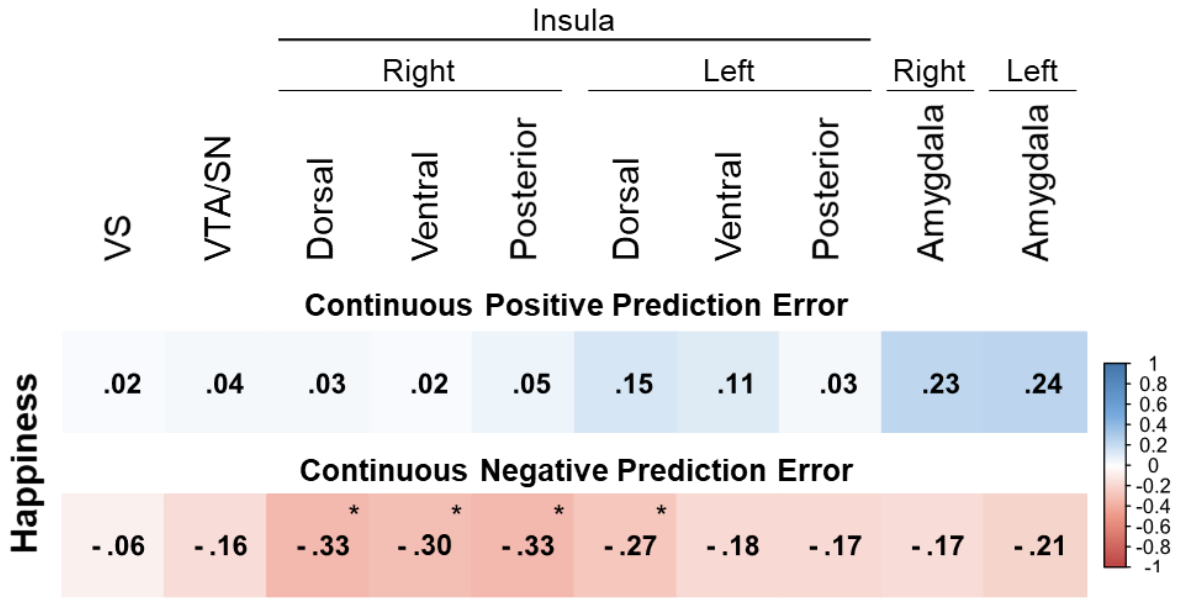


Supplementary Figure 4.5. Spearman correlation between learning rates for self-related positive (PE+) and negative prediction errors (PE-) and brain activity within ROIs in response to PE+ and PE- (categorical PE effect). * $p < .05$, x FDR corrected.

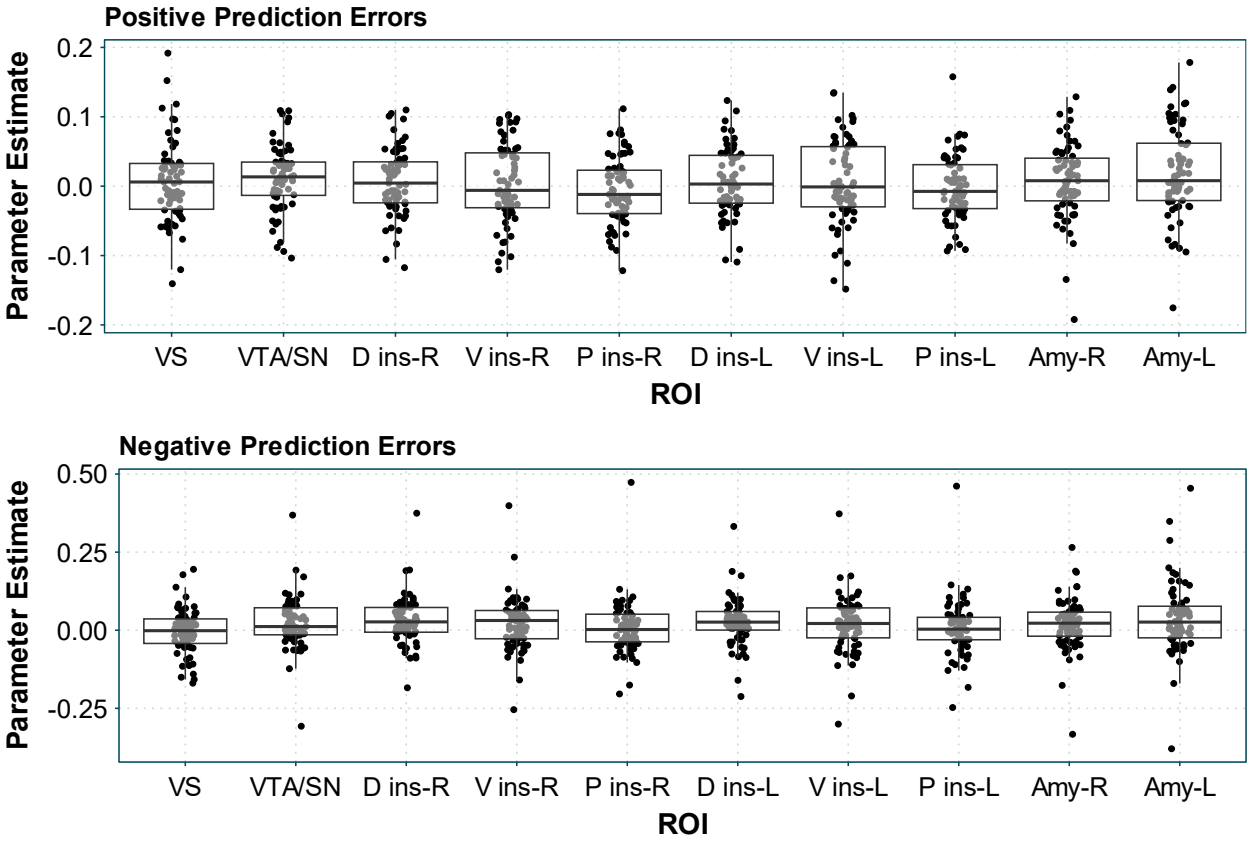


Supplementary Figure 4.6. Neural activation associated with negative compared to positive prediction error processing (continuous PE Valence effect). a) Processing of self-related negative prediction errors (PE-) relative to positive prediction errors PE+ was associated with increased activation of the superior frontal gyrus, paracingulate/ cingulate gyrus, frontal orbital cortex, operculum/ insula cortex ($p < .05$, FWE corrected at peak level for the whole brain). Anatomical labels were derived from the SPM Anatomy Toolbox Version 3.0.

Association of happiness and brain activity for positive and negative prediction errors

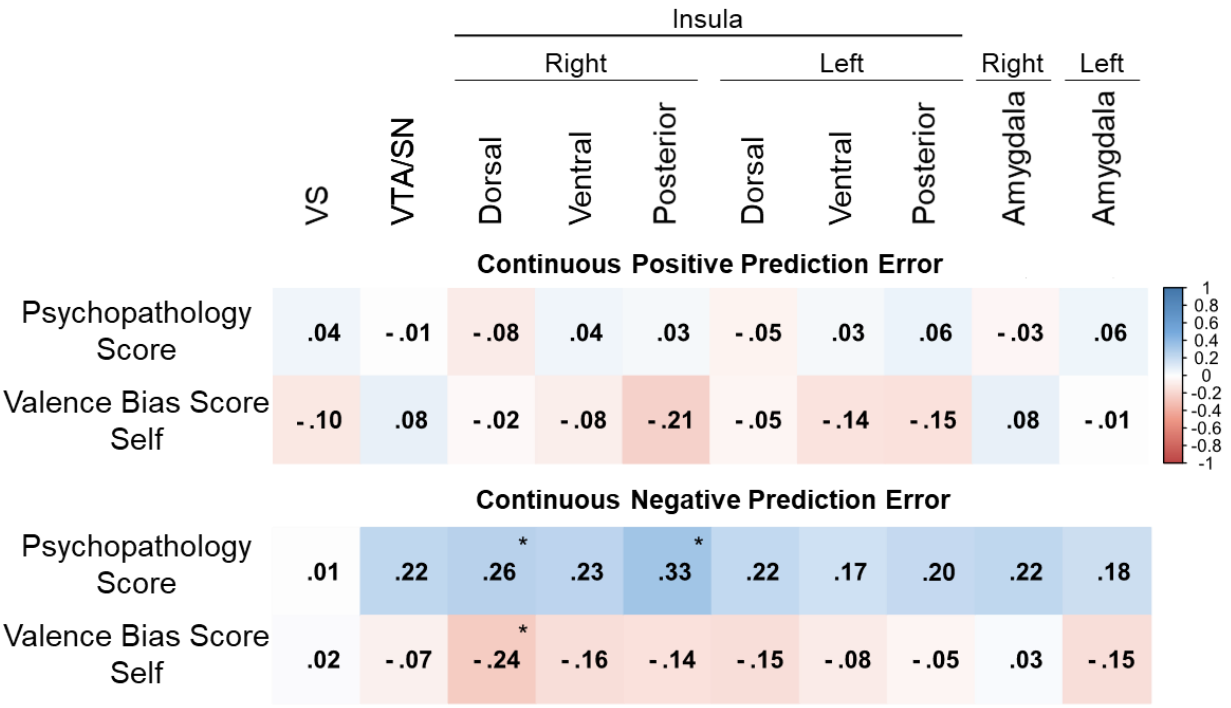


Supplementary Figure 4.7. Spearman correlation between the happiness ratings and brain activity within ROIs in response to more positive and negative prediction errors (continuous PE effect). * $p < .05$.

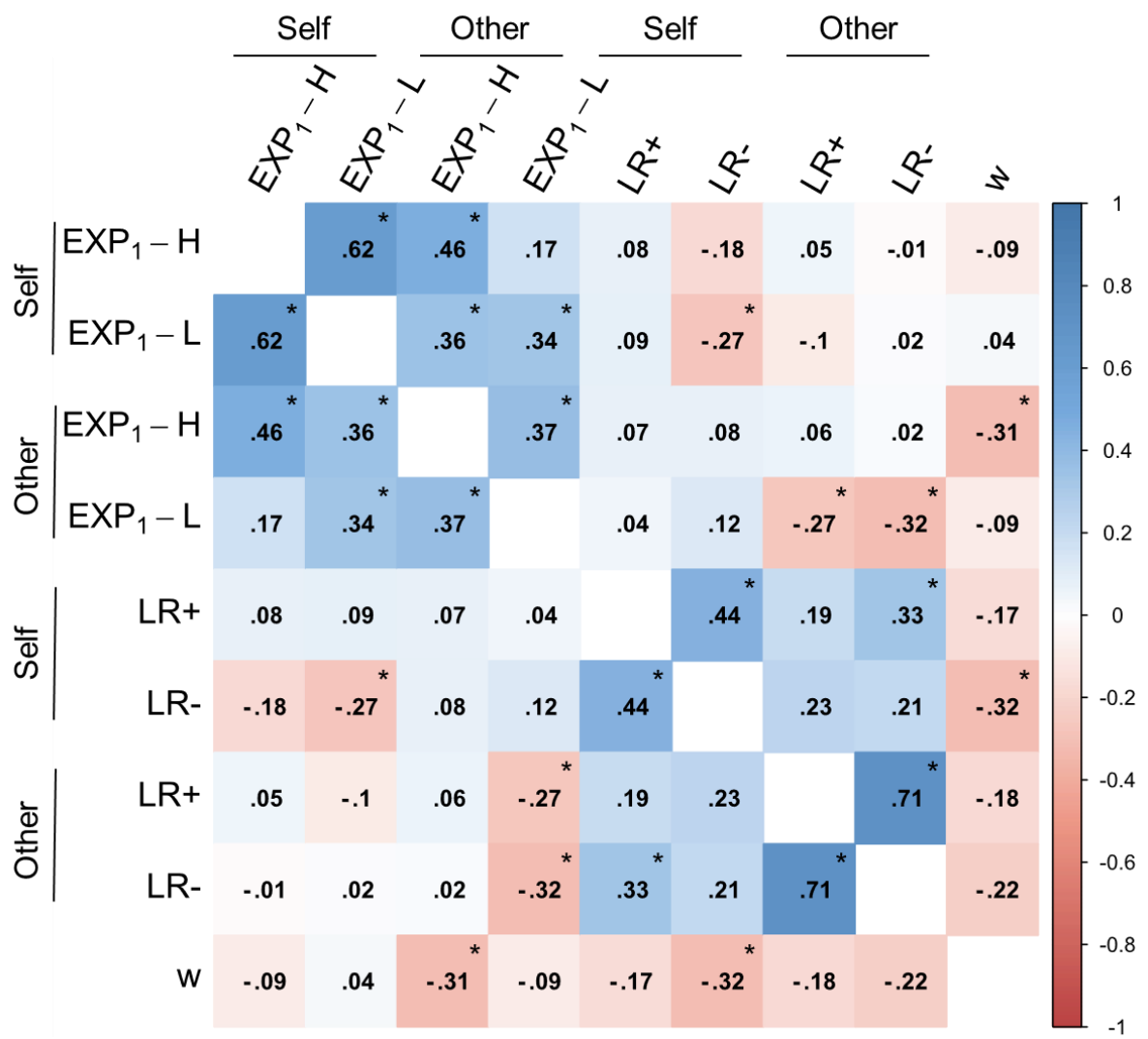


Supplementary Figure 4.8. Parameter estimates for the processing of positive and negative prediction errors *(continuous PE Valence effect)* in the regions of interest (ROIs: VS = ventral striatum, ins = insula cortex, Amy = amygdala, D = dorsal, V = ventral, P = posterior, R = right, L = left). Parameter estimates were derived from all voxels within each ROI and averaged across all voxels. Dots = data point from individual subject, center line of the box = median, edge lines of the box = first/ third quartile, Whiskers = largest/ smallest data point at most 1.5 times the interquartile range above/ below the respective border.

Association of psychopathology and self-related learning bias with brain activity for positive and negative prediction errors



Supplementary Figure 4.9. Spearman correlation between the Valence Bias Score VBS = (LR+ - LR-)/ (LR+ + LR-), as well as the Psychopathology score and brain activity within ROIs in response to more positive and negative prediction errors *(*continuous PE effect). * $p < .05$.



Supplementary Figure 4.10. Spearman correlation of model parameters for the winning model. EXP_1 = estimated first expected performance for the two ability conditions (H = high, L = low), separately for Self and Other. LR = learning rates for positive (LR+) and negative prediction errors (LR-), separately for Self and Other. w = weighting factor. * $p < .05$.

Supplementary Tables

Supplementary Table 4.1. Sample characteristics - group comparison

	CON		MDD		*t*-test		
	M	*SD*	*M*	*SD*	*t*(65)	*p*	Cohen's *d*
Age	34.25	10.23	34.23	10.54	0.01	0.995	0
BDI sum	13.88	8.00	53.91	12.00	-15.92	<.001	-3.85
SIAS mean	0.90	0.41	1.98	0.73	-7.38	<.001	-1.78
ATQ mean	1.87	0.29	3.49	0.55	-14.84	<.001	-3.59

Note. Group comparison of sample characteristics using a two-sample *t*-test. Beck's depression inventory (BDI-V, Schmitt et al., 2003, items on a 6-point Likert scale from zero to five), Automatic Thought Questionnaire (ATQ, Pössel et al., 2005, items on a 5-point Likert scale from one to five), Social Interaction Anxiety scale (SIAS, Stangier et al., 1999, items on a 5-point Likert scale from zero to four). *M* = mean, *SD* = standard deviation, *d* = Cohen's *d* with Hedges correction. MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary Table 4.2. Model-free analysis of belief updating behavior in individuals with and without depression

	Estimate	*SE*	95% CI		*df*	*t*	*p*
			lower	upper			
(Intercept)	51.94	1.35	49.31	54.58	5279	38.61	< .001
Ability [High]	5.49	1.25	3.04	7.94	5279	4.39	< .001
Agent [Self]	-3.62	1.25	-6.06	-1.17	5279	-2.89	.004
Trial	-0.56	0.07	-0.70	-0.41	5279	-7.54	< .001
Group [MDD]	-1.19	1.86	-4.90	2.53	65	-0.64	.526
Ability x Agent	-0.81	1.77	-4.27	2.65	5279	-0.46	.649
Ability x Trial	1.08	0.10	0.88	1.28	5279	10.34	< .001
Ability x Group	-0.86	1.73	-4.24	2.53	5279	-0.49	.621
Agent x Trial	-0.27	0.10	-0.47	-0.07	5279	-2.59	.010
Agent x Group	-1.90	1.73	-5.29	1.48	5279	-1.10	.271
Trial x Group	0.09	0.10	-0.11	0.29	5279	0.85	.394
Ability x Agent x Trial	0.07	0.15	-0.22	0.36	5279	0.45	.649
Ability x Agent x Group	0.36	2.45	-4.42	5.15	5279	0.15	.882
Ability x Trial x Group	-0.08	0.14	-0.36	0.21	5279	-0.54	.593
Agent x Trial x Group	-0.03	0.14	-0.32	0.25	5279	-0.23	.819
Ability x Agent x Trial x Group	0.08	0.20	-0.32	0.48	5279	0.39	.697

Note. Trial (continuous, 1-20) x Agent [Self, Other] x Ability condition [High, Low] x Group [MDD, CON] linear mixed-effects model fit by maximum likelihood. Dependent variable: trial-by-trial ratings of expected performance. Beta-estimate, *SE* = standard error, *CI* = confidence interval, *df* = degrees of freedom, *t* = t-value, MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary Table 4.3. Model Space

Assumptions about learning	Model	Learning Parameters
Self = Other		
No learning bias	Unity Model (M1)	***α***
Prediction Error: Positive ≠ Negative	Valence Model (M2)	***α***PE+, ***α***PE-
Ability Contex: High ≠ Low	Ability Model (M3)	***α***High, ***α***Low
Self ≠ Other		
No learning bias	Unity Model (M4)	***α***Self ***α***Other
Prediction Error: Positive ≠ Negative	Valence Model (M5)	***α***Self/PE+, ***α***Self/PE- ***α***Other/PE+, ***α***Other/PE-
Prediction Error: Positive ≠ Negative & Decay for extreme feedback values	Weighted Valence Model (M6)	***α***Self/PE+, ***α***Self/PE- ***α***Other/PE+, ***α***Other/PE- ***w***
Ability Contex: High ≠ Low	Ability Model (M7)	***α***Self/High, ***α***Self/Low ***α***Other/High, ***α***Other/Low
No learning	Mean Model (M0)	

Note: Four starting values (estimated first expectation) for the four feedback conditions (Agent [Self vs. Other] x Ability [High Ability vs. Low Ability] were additionally estimated for all models.

Supplementary Table 4.4. Model comparison

Model		*PSIS-LOO*	*LOO-SE*	*LOO-Diff*	*% of* $\hat{k} > 0.7$	*No. Est. Parameters*
Whole sample						
	Mean Model (M0)	-2441.5	216.15	-1176.0	0.04	4
Self = Other	Unity Model (M1)	-1722.6	216.37	-457.1	0.62	5
	Valence Model (M2)	-1580.0	210.76	-314.5	0.34	6
	Ability Mode (M3)	-1639.2	213.23	-373.7	0.71	6
Self ≠ Other	Unity Model (M4)	-1644.7	209.58	-379.2	0.49	6
	Valence Model (M5)	-1356.1	219.24	-90.6	0.39	8
	Weighted Valence Model (M6)	-1265.5	229.54	-	1.10	9
	Ability Model (M7)	-1539.0	208.47	-273.4	1.32	8
MDD						
	Mean Model (M0)	-1404.3	156.32	-603.8	0.02	4
Self = Other	Unity Model (M1)	-1089.8	182.84	-289.3	0.49	5
	Valence Model (M2)	-1000.3	176.65	-199.8	0.19	6
	Ability Model (M3)	-1039.4	179.52	-238.9	0.58	6
Self ≠ Other	Unity Model (M4)	-1022.5	175.15	-222.0	0.22	6
	Valence Model (M5)	-835.4	182.05	-34.9	0.17	8
	Weighted Valence Model (M6)	-800.5	191.68	-	0.58	9
	Ability Model (M7)	-961.9	173.61	-161.4	0.76	8
CON						
	Mean Model (M0)	-1037.2	148.25	-572.1	0.02	4
Self = Other	Unity Model (M1)	-632.8	108.87	-167.8	0.13	5
	Valence Model (M2)	-579.8	109.48	-114.7	0.15	6
	Ability Model (M3)	-599.9	108.84	-134.8	0.13	6
Self ≠ Other	Unity Model (M4)	-622.2	110.69	-157.2	0.26	6
	Valence Model (M5)	-520.8	121.02	-55.7	0.22	8
	Weighted Valence Model (M6)	-465.0	124.64	-	0.52	9
	Ability Model (M7)	-577.0	111.48	-112.0	0.56	8

Note. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (weighted Valence Model) and standard errors of differences; percentage of $\hat{k}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al.); No. Est. Parameters = number of estimated parameters in the model; S = O: same learning rates for Self and Other, S ≠ O: separate learning rates for Self and Other; MDD = Major depressive disorder (n = 35), CON = control group (n = 32).

Supplementary Table 4.5. Model-based analysis of belief updating behavior (learning rates) in individuals with and without depression

	Estimate	*SE*	95% CI		*df*	*t*	*p*
			lower	upper			
(Intercept)	0.22	0.03	0.17	0.27	195	8.83	< .001
PE Val [PE+]	0.04	0.03	-0.02	0.10	195	1.29	.199
Agent [Self]	0.07	0.03	0.01	0.13	195	2.35	.020
Group	-0.01	0.04	-0.08	0.06	65	-0.30	.764
PE Val x Agent	-0.14	0.04	-0.23	-0.06	195	-3.29	.001
PE Val x Group	-0.01	0.04	-0.09	0.07	195	-0.23	.818
Agent x Group	0.03	0.04	-0.06	0.11	195	0.62	.533
PE Val x Agent x Group	0.01	0.06	-0.11	0.13	195	0.18	.861

Note. Prediction Error Valence (PE Val, [PE+, PE-]) x Agent [Self, Other] x Group [MDD, CON] linear mixed-effects model fit by maximum likelihood. Dependent variable: learning parameters of the winning model. Beta-estimate, *SE* = standard error, *CI* = confidence interval, *df* = degrees of freedom, *t* = *t*-value, MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary Table 4.6. Increasingly biased updating for self and other is linked to levels of psychopathology

	Estimate	*SE*	95% CI		*df*	*t*	*p*
			lower	upper			
(Intercept)	0.07	0.03	0.00	0.14	65	2.11	.039
Agent [Self]	-0.24	0.05	-0.34	-0.15	65	-4.93	< .001
Psychopathology	0.06	0.03	-0.01	0.13	65	1.64	.106
Agent [Self] x Psychopathology	-0.10	0.05	-0.2.	0.00	65	-2.03	.047

Note. Agent [Self, Other] x Psychopathology score linear mixed-effects model fit by maximum likelihood across the whole sample. Dependent variable: Valence Bias Score = (LR+ - LR-)/ (LR+ + LR-). Beta-estimate, *SE* = standard error, *CI* = confidence interval, *df* = degrees of freedom, *t* = *t*-value, sample size: *n* = 67.

Supplementary Table 4.7. Baseline activations associated with the tracking of positive and negative prediction errors

Contrasts/ Brain regions	Side	Cluster Size	MNI Coordinates			*T*	*p*
			x	y	z		
Positive prediction error							
Angular Gyrus/ Lateral Occipital Cortex, superior division	R	69	-56	-60	24	5.40	.012
Angular Gyrus/ Supramarginal Gyrus, posterior division			-46	-50	24	5.20	.022
Angular Gyrus/ Lateral Occipital Cortex, superior division	L	8	52	-54	24	5.15	.026
Negative prediction error							
Superior Frontal Gyrus/ Juxtapositional Lobule Cortex	R/ L	245	-2	12	60	7.29	<.001
Inferior Frontal Gyrus, pars triangularis/ pars opercularis	L	145	48	28	4	6.34	<.001
			50	20	2	5.45	.009
Frontal Orbital Cortex/ Frontal Operculum Cortex			40	28	-4	5.30	.014
Superior Frontal Gyrus/ Paracingulate Gyrus		65	-6	52	22	5.59	.005
Paracingulate Gyrus/ Superior Frontal Gyrus			8	50	22	5.48	.008
Frontal Orbital Cortex/ Frontal Operculum Cortex		19	-38	26	-6	5.23	.017

Note. Baseline activations of prediction error tracking. Positive and negative prediction error refer to the unsigned prediction error values as two parametric modulators for the feedback phase of the Self condition. The *p*-values are FWE-corrected for the whole brain at peak level. Whole sample: $n = 67$. *R* = right, *L* = left. *T* = *t*-value. Anatomical labels were derived from the SPM Anatomy Toolbox Version 3.0.

Supplementary Table 4.8. Group comparison of positive and negative prediction error tracking within regions of interest

Covariates/ Regions of interest	Side	Cluster Size	MNI Coordinates			*T*	*p*
			x	y	z		
Positive prediction error - CON > MDD							
All ROIs	No suprathreshold clusters						
Negative prediction error - MDD > CON							
Dorsal insula - right	R	237	36	16	-8	3.24	.030
Dorsal insula - left	L	132	-40	16	-10	2.67	.088
Ventral insula - right	R	389	36	-14	2	3.10	.040
Ventral insula - left	L	6	-38	12	-10	2.20	.261
Posterior insula - right	R	272	40	-18	12	3.30	.014
Posterior insula - left	L	33	-36	-22	14	2.39	.166
Amygdala - right	R	241	34	0	-12	2.48	.129
Amygdala - left	L	107	-24	-2	-18	2.58	.108
Ventral Striatum (functional PE ROI)	No suprathreshold clusters						
VTA/SN	L	915	-22	-18	-8	3.36	.046
Interaction Group x PE Valence MDD (PE-), CON (PE+) > MDD (PE+), CON (PE-)							
Dorsal insula - right	R	123	36	20	-6	2.90	.070
Dorsal insula - left	L	24	-38	14	-10	2.43	.187
Ventral insula - right	R	85	38	-12	2	3.35	.029
Ventral insula - left	L	3	-30	6	-14	2.40	.244
Posterior insula - right	R	175	36	-18	4	3.17	.030
Posterior insula - left	L	3	-30	-26	18	2.03	.328
Amygdala - right	R	58	34	2	-16	2.27	.236
Amygdala - left	L	39	-26	-6	-18	2.75	.086
Ventral Striatum (functional PE ROI)	No suprathreshold clusters						
VTA/SN	L	123	-20	-14	-6	2.91	.185

Note. Group comparison of prediction error tracking for positive (PE+) and negative (PE-) prediction errors using two-sample *t*-tests and Group x PE Valence interaction using a flexible factorial design. The *p*-values are FWE corrected within ROIs at peak level. All three tests had no suprathreshold clusters on the whole brain level (FWE-corrected). *R* = right, *L* = left. *T* = *t*-value. MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary Table 4.9. Emotions - group comparison

	CON		MDD		*t*-test		
	M	*SD*	*M*	*SD*	*t*(65)	*p*	*d*
Happiness	58.28	12.04	42.98	21.95	3.58	< .001	0.85
Arousal	30.41	20.89	41.70	23.91	-2.06	.043	-0.50
Pride	45.06	19.66	34.73	22.62	2.00	.050	0.48
Embarrassment	20.34	18.61	23.48	23.91	-0.60	.549	-0.14
Tiredness	34.47	22.41	44.56	26.32	-1.69	.095	-0.41

Note. Group comparison of emotion ratings during task performance. Two ratings per emotion on a continuous scale from zero to 100 following the presentation of self-related feedback trials. *M* = mean, *SD* = standard deviation, *d* = Cohen's *d* with Hedges correction. MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary Table 4.10. Prior beliefs about (estimation) abilities from global to specific - group comparison

	CON		MDD		*t*-test		
	M	*SD*	*M*	*SD*	*t*(65)	*p*	*d*
General abilities (SDQ-III)	6.66	0.56	4.50	1.44	8.23	< .001	1.93
General estimation ability	5.09	1.00	4.49	1.36	2.10	.040	0.50
Specific estimation ability - self	41.14	18.41	36.51	17.81	1.04	.301	0.25
Confidence of specific estimation ability - self	26.44	12.23	25.03	12.13	0.47	.638	0.11
Specific estimation ability - other	49.12	14.64	46.21	14.38	0.82	.415	0.20
Confidence of specific estimation ability - other	34.79	14.97	25.46	13.99	2.63	.011	0.64
Ability expectation rating 1. trial - self	52.06	7.27	47.57	12.36	1.83	.072	0.43
Ability expectation rating 1. trial - other	57.27	6.14	54.61	9.30	1.39	.170	0.33

Note. Group comparison of self-beliefs and beliefs about the other person prior to the task. 1. General ability/ self-esteem measured with the Self-Description Questionnaire-III (SDQ-III, subscale, Marsh & O'neill, 1984), 2. General estimation ability (one item): 'I'm good at estimating,' 8-point Likert scale, 3./5. Specific estimation ability for self and other (one item per estimation category): 'How well do you think [you are/ the other person is] at estimating [e.g., the weight of animals]?', 4./6. Confidence: 'How sure are you that your assessment of [your own/ the other person's] ability is correct?', 7./8. First rating of the main task before the first feedback presentation. *M* = mean, *SD* = standard deviation, *d* = Cohen's *d* with Hedges correction. MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary Table 4.11. Setting characteristics at the start - group comparison

	CON		MDD		*t*-test		
	M	*SD*	*M*	*SD*	*t*(65)	*p*	*d*
Category-specific estimation experience	2.44	1.32	3.01	1.27	-1.82	.073	-0.44
Importance of estimation ability	2.80	1.39	3.17	1.40	-1.10	.277	-0.26
Familiarity of the other person	1.09	0.3	1.20	0.41	-1.23	.223	-0.29
Likeability of other person	5.16	1.14	5.20	1.23	-0.15	.880	-0.04
Estimated own likeability rated by other	4.69	1.06	4.09	1.60	1.83	.072	0.43

Note. Group differences in previous experience, importance of estimating, and attitude towards the other person. All characteristics rated on one item each on a 7-point Likert scale in a pre-survey prior to the main task. Item wordings: 1. 'How much experience do you have with estimating [e.g., the weight of animals]?', 2. 'How important is it to you to be good at estimating [e.g., the weight of animals]?', 3. 'How well do you know the other person?', 4. 'I like the other person.', 5. 'I think the other person finds me likable.'. *M* = mean, *SD* = standard deviation, *d* = Cohen's *d* with Hedges correction. MDD = Major depressive disorder (*n* = 35), CON = control group (*n* = 32).

Supplementary References

Marsh, H. W., & O'neill, R. (1984). Self Description Questionnaire iii: The construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement*, *21*(2), 153–174.

Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Schröder, A., Stolz, D. S., Paulus, F. M., & Krach, S. (2022). Neurocomputational mechanisms of affected beliefs. *Communications Biology*, *5*(1), 1241.

Pössel, P., Seemann, S., & Hautzinger, M. (2005). Evaluation eines deutschsprachigen Instrumentes zur Erfassung positiver und negativer automatischer Gedanken [Evaluation of a German-language instrument for recording positive and negative automatic thoughts]. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *34*(1), 27–34.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*, 971–985.

Schmitt, M., Beckmann, M., Dusi, D., Maes, J., Schiller, A., & Schonauer, K. (2003). Messgüte des vereinfachten Beck-Depressions-Inventars (BDI-V) [Measurement quality of the simplified Beck Depression Inventory (BDI-V)]. *Diagnostica*, *49*(4), 147–156.

Stangier, U., Heidenreich, T., Berardi, A., Golbs, U., & Hoyer, J. (1999). Die Erfassung sozialer Phobie durch die Social Interaction Anxiety Scale (SIAS) und die Social Phobia Scale (SPS) [The assessment of social phobia using the Social interaction anxiety scale (SIAS) and the social phobia scale (SPS)]. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *28*(1), 28–36.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research: JMLR, 17*, 3581–3618.

5 Study 4

Self-beneficial belief updating as a coping mechanism for stress-induced negative affect[4]

5.1 Abstract

Being confronted with social-evaluative stress elicits a physiological and a psychological stress response. This calls for regulatory processes to manage negative affect and maintain self-related optimistic beliefs. The aim of the current study was to investigate the affect-regulating potential of self-related updating of ability beliefs after exposure to social-evaluative stress, in comparison to non-social physical stress or no stress. We assessed self-related belief updating using trial-by-trial performance feedback and described the updating behavior in a mechanistic way using computational modeling. We found that social-evaluative stress was accompanied by an increase in cortisol and negative affect which was related to a positive shift in self-related belief updating. This self-beneficial belief updating, which was absent after physical stress or control, was associated with a better recovery from stress-induced negative affect. This indicates that enhanced integration of positive self-related feedback can act as a coping strategy to deal with social-evaluative stress.

[4] This study has been published as: **Czekalla, N.**, Stierand, J., Stolz, D. S., Mayer, A. V., Voges, J. F., Rademacher, L., Paulus, F. M., Krach, S., & Müller-Pinzler, L. (2021). Self-beneficial belief updating as a coping mechanism for stress-induced negative affect. *Scientific Reports*, *11*(1), 1–13.
My contribution: designing the research, data acquisition, data analysis, and writing the manuscript.

5.2 Introduction

Human beings strive to be accepted by others and to maintain a positive social image (Baumeister & Leary, 1995). Thus, social evaluation of our behavior can pose a threat to our social image, eliciting a stress response in our body (Burke, 1991; Kirschbaum et al., 1993a; Rohleder et al., 2007). This initiates various physiological processes (Joëls & Baram, 2009) and is associated with negative affective consequences, like anxiety or embarrassment (Campbell & Ehlert, 2012; Gruenewald et al., 2004; Müller-Pinzler et al., 2015). Social evaluation, however, is fundamental to self-related learning processes, as it gives one the opportunity to integrate the feedback we receive from others and update the beliefs about ourselves accordingly (Eisenberger et al., 2011; Markus & Wurf, 1987). Biases in how we process self-related feedback on our behaviors, i.e. whether we focus more on negative or positive feedback, impact our affective reactions (Gotlib & Krasnoperova, 1998; Roese & Olson, 2007) and, in the case of self-serving processing, may function as a coping strategy (Roese & Olson, 2007). While (social) stress is a risk factor for many psychiatric conditions (Kessler et al., 1985), successful coping is an important factor in maintaining mental health (Gloria & Steinhardt, 2016). In the current study we implemented a computational modeling approach to investigate the coping mechanism of self-beneficial belief updating after social-evaluative stress and tested whether shifted information processing after stress predicts recovery from stress-induced negative affect.

When we receive feedback regarding our behaviors, information processing and belief updating is shaped by self-relevant motivations (Bromberg-Martin & Sharot, 2020), especially the motivation to maintain optimistic beliefs about the self (Sharot & Garrett, 2016). Many studies have demonstrated that the process of self-related belief updating is biased in favor of positive information, i.e. self-related beliefs are updated more strongly when feedback is better than expected (Eil & Rao, 2010; Korn et al., 2012; Mobius et al., 2011; Sharot et al., 2011). However, updating biases towards negative feedback have been reported in performance contexts (Ertac, 2011; Müller-Pinzler et al., 2019), which indicates that the context of learning (i.e. learning about own abilities or learning about one's personality), type of feedback and prior assumptions are important factors when explaining self-related belief updating biases.

While there are only relatively few studies on the effects of stress on self-related belief updating, various studies on reward processing and non-self-related feedback processing have shown that stress is an influencing factor in this regard. One key mechanism for feedback-based learning is the prediction error signal, indicating the difference between a predicted and an actual outcome (Glimcher, 2011; Watabe-Uchida et al., 2017), which is being minimized by updating beliefs during learning. This signal is generated by dopaminergic neurons of the ventral striatum (Schultz et al., 1997), which might be particularly important for the stress-induced modulation of prediction error signals as the dopamine system is sensitive to stress (Adler et al., 2000; Payer et al., 2017). However, these effects depend on the type, intensity and schedule of the stress exposure (Holly & Miczek, 2016), which might also explain heterogeneous effects of stress on reward

processing and feedback-based learning. Research on declarative memory has shown that timing of stress matters. In the acute stress phase, mainly characterized by a rapid sympathetic response, catecholamines and non-genomic glucocorticoid actions lead to increased memory formation of the stressful event. Cortisol is released with a delay and inhibits memory consolidation later on to avoid interference with non-stress-related information (Joëls et al., 2006; Schwabe et al., 2012). Besides this inhibition of hippocampus-dependent declarative memory, neuro-imaging research on classification learning also found a shift towards striatum-based procedural learning after stress, i.e. also non-declarative learning processes are modulated by stress (Schwabe & Wolf, 2012). Acute stress is associated with an increased extinction resistance in fear conditioning (Antov et al., 2013). When learning takes place with a delay to stress, trace conditioning (Wolf et al., 2009), and updating in reversal fear conditioning (Raio et al., 2017) are attenuated, cortisol is associated with reduced fear conditioning (Antov et al., 2013) and working memory is reduced (Luethi et al., 2008; Qin et al., 2009). Timing of stress seems to be important for feedback-based or reward-based learning as well (van Leeuwen et al., 2019a). Initially, acute stress (e.g. a threat of a shock during learning) impairs feedback-based learning of reward (Bogdan & Pizzagalli, 2006). Neurally, acute stress attenuates the response to reward in the striatum and orbitofrontal cortex (Kumar et al., 2014; Porcelli et al., 2012) and enhances the striatal response to aversive feedback (Robinson et al., 2013). Accordingly, under acute stress self-related belief updating is more strongly driven by unfavorable feedback, i.e. the learning bias in favor of positive information (optimism bias) usually found in self-related belief updating is absent (Garrett et al., 2018b). The opposite effects are reported when learning takes place with a delay to stress (e.g. after a public speech), a phase mainly characterized by an increase of cortisol (Schwabe et al., 2012). Here, feedback processing is more strongly driven by stimuli signaling reward and possibly associated with stress-induced cortisol change (Lighthall et al., 2013) while learning from negative feedback is decreased, potentially linked to cortisol levels before learning (Petzold et al., 2010). On the neural systems level, stress recovery is associated with increased striatal responses to rewarding feedback at 50 min after stress (van Leeuwen et al., 2019a, 2019b). Moreover, specifically individuals with low striatal reward reactivity showed an association of recent life stress with lower positive affect, which makes striatal reactivity a potential factor of successful stress coping (Nikolova et al., 2012).

According to classic appraisal theories of stress (Lazarus & Folkman, 1984), different strategies such as seeking social support, positive revaluation or acceptance are helpful in coping with stress-induced negative affect (Glanz & Schwartz, 2008; Lazarus & Folkman, 1984; Thoits, 1995). In the context of social-evaluative stress a self-protection strategy is to view oneself in a positive light, i.e. emphasizing the own desirability, focusing on own successes and attributing failure externally (vanDellen et al., 2011). This strategy has also been successful in alleviating stress-induced negative affect following a performance situation (Jundt & Hinsz, 2002; Roese & Olson, 2007). Generally, an optimistic way of processing self-related feedback has been associated with better mental

health (Sharot, 2011; Taylor & Brown, 1988). On the contrary, processing self-related feedback in a more negative way may result in negative beliefs about the self (Beck, 2002) and ultimately lead to lower self-esteem or depressive symptomatology. Studies on self-related belief updating in individuals with depression suggest that information processing is distorted in a negative direction (Korn et al., 2014) and that coping strategies for situations of social-evaluative stress are less readily available in these patients (Greenberg et al., 1992).

In the present study, we aim to investigate the effects of social-evaluative stress on the updating of self-related ability beliefs and the propensity to engage into self-beneficial learning after social-evaluative stress. By means of two well validated and highly reliable paradigms, the Trier Social Stress Test (public speech; Kirschbaum et al., 1993a) and the Cold Pressor Test (Hines & Brown, 1933), as well as a no stress control condition, we directly manipulated levels of social-evaluative stress in a between-groups design. After stress manipulation we used computational modeling to describe participants' self-related belief updating behavior using the learning of own performance (LOOP) task (Müller-Pinzler et al., 2019). In this task participants continuously update beliefs about their abilities in epistemologically novel behavioral domains. We then used participants' learning bias from positive and negative feedback to predict their recovery from stress-induced negative affect. We found that social but not physical stress shifted subsequent self-related belief updating in a more self-beneficial direction which predicted better recovery from negative affect. We elaborate on the relationship between stress (specifically the components of negative affect and cortisol response), self-related belief updating and affect regulation in healthy participants and discuss the potential of our findings for a better understanding of maladaptive self-related belief systems in psychiatric conditions such as depression.

5.3 Results

After exposure to social-evaluative stress (SOC, Trier Social Stress Test), non-social, physical stress (PHY, Cold Pressor Test) or a no stress control condition (CON, reading) participants performed the LOOP task (Müller-Pinzler et al., 2019), which was covered as a measure of cognitive estimation skills (see Figure 5.1). The central idea of the LOOP task is to create a performance context and provide manipulated positive or negative feedback in comparatively neutral domains in which people have only vague prior assumptions. By this means, individuals form a concept about their own abilities over the course of the experiment. In a previous study, we showed that this process of self-related belief updating can be described best by a computational prediction error learning model (adapted from Rescorla and Wagner, 1972) with two separate learning parameters for positive and negative prediction errors (Müller-Pinzler et al., 2019). During the LOOP task, participants were asked to answer estimation questions in two different estimation domains (e.g. estimating the weight of animals and the height of buildings) and received manipulated performance feedback implying a rather good performance in one category and a rather bad performance in the other one (high vs. low ability condition). In the beginning of each trial participants saw a cue indicating the estimation category and had

to rate their expected performance for the upcoming estimation question in this category. A manipulated feedback on their estimation performance in relation to an alleged reference group was presented afterwards. Saliva cortisol as well as negative affect, including perceived stress, embarrassment, anger, and frustration, were assessed several times during the experiment. Pre-stress baseline measures ($T1_{AFF/CORT}$) were taken after a 10-min period of rest in the beginning of the session. Post-stress negative affect was rated immediately after the stress exposure or control task ($T2_{AFF}$) to calculate the mean change of negative affect (ΔAFF). Post-stress cortisol samples were taken after another 10-min period of rest ($T2_{CORT}$) to calculate the mean cortisol change (ΔCORT). After performing the LOOP task, saliva samples and negative affect were again obtained ($T3_{AFF/CORT}$, for a detailed description see methods).

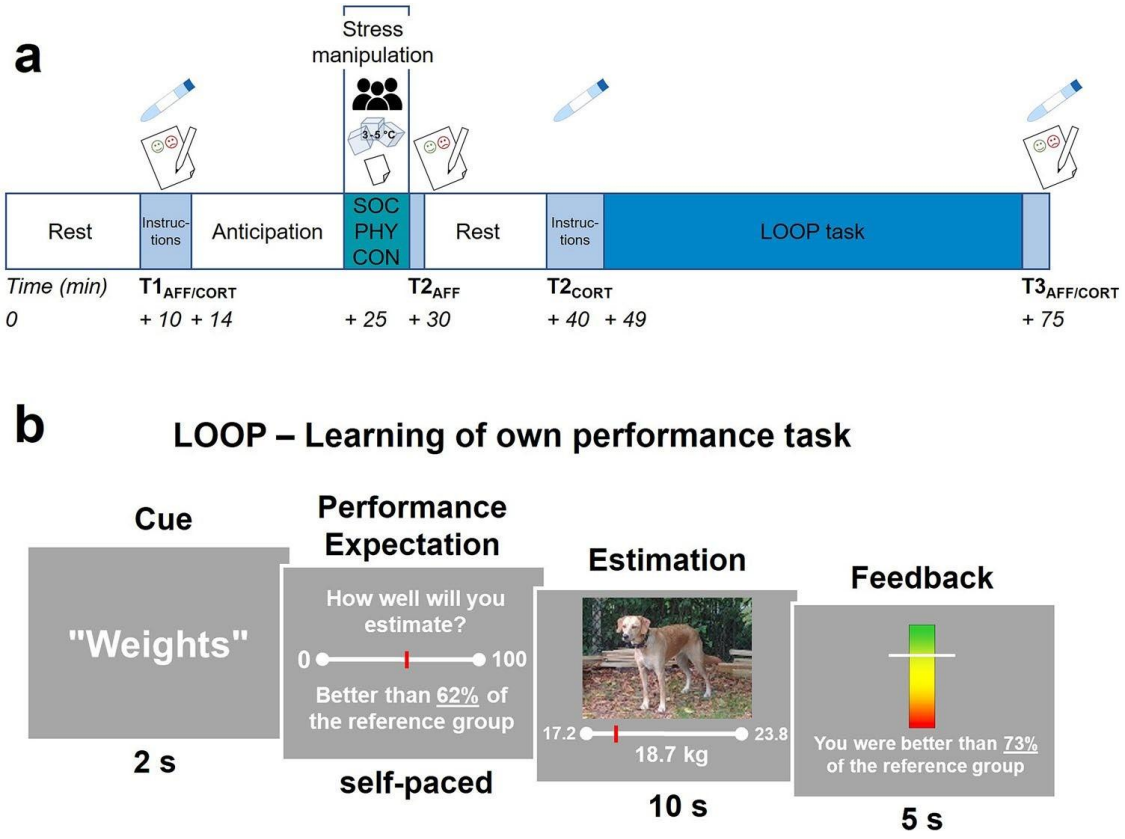


Figure 5.1. **(a)** Experimental timeline and procedure. SOC: social-evaluative stress group (public speech [audience icon], *n* = 29), PHY: physical stress group (Cold Pressor Test [ice cubes icon], *n* = 30), CON: no stress control group (reading task [paper icon], *n* = 30), salivette icon: saliva collection for cortisol determination; paper pencil icon: rating of negative affect including perceived stress, embarrassment, anger, and frustration. **(b)** Sequence of one trial. 1. Cue: display of the upcoming estimation category associated with a high or low ability condition, 2. Performance expectation rating, 3. Estimation question, 4. Performance feedback. Figure adapted from Müller-Pinzler et al. (2019).

Cortisol response and negative affect. *Cortisol change.* The stress manipulation was effective and social-evaluative stress, as well as physical stress, led to a stronger increases of cortisol levels from baseline $T1_{CORT}$ to post-stress $T2_{CORT}$ than in the no stress control group (Scheirer-Ray-Hare test on ΔCORT controlled for time of the day [TIME]: main effect factor Stress group $H_2 = 18.9$, $p < .001$, post-hoc Dunn-Bonferroni-Tests for

factor Stress group: SOC vs. CON: z = -4.29, $p < .001$; PHY vs. CON: z = -2.76, $p = .018$; see Supplementary Table 5.1). There was no statistically significant difference between the two stress groups (SOC vs. PHY: z = 1.56, $p = .355$; baseline cortisol levels did not significantly differ between groups $H_2 = 1.74$, $p = .419$ controlled for TIME, see Figure 5.2a and c).

Change in negative affect. Mean negative affect increased significantly after social-evaluative stress but not after physical stress compared to the control group (Kruskal Wallis test on ΔAFF: $H_2 = 43.9$, $p < .001$, post-hoc Dunn-Bonferroni-tests: SOC vs. CON: $z = -6.45$, $p < .001$, PHY vs. CON: $z = -1.88$, $p = .182$, SOC vs. PHY: $z = -4.59$, $p < .001$; baseline negative affect did not significantly differ between groups $H_2 = 3.2$, $p = .201$; see Figure 5.2b and d and Supplementary Figure 5.1).

Forming self-related beliefs over time. In a model free behavior analysis we replicated previous findings regarding the LOOP task which indicates self-related belief updating in response to the feedback (Müller-Pinzler et al., 2019). Over the time of 30 trials, participants adapted their performance expectation ratings (EXP) towards the positive and negative feedback of the two ability conditions, i.e. they updated their self-related beliefs (Figure 5.3c, significant factor Ability condition high vs. low $t_{86} = 8.52$, $p < .001$, significant Trial x Ability condition interaction $t_{5156} = 32.72$, $p < .001$). Social-evaluative stress modulated self-related belief updating over time, i.e. performance expectation ratings became increasingly higher compared to physical stress or no stress (Trial x Ability condition x Stress group split into the contrasts social [SOC] vs. non-social [PHY, CON] and the orthogonal contrast PHY vs. CON: interaction for contrast SOC vs. [PHY, CON]; $t_{5156} = 4.01$, $p < .001$). In the physical stress group performance expectation ratings were even more negative over time than in the no stress control condition (Trial x Ability condition x Contrast PHY vs. CON $t_{5156} = -2.15$, $p = .031$; mixed-effects model with the within-group factor Ability condition, the between-group factor Stress group, and the continuous variable trial, plus interactions, see Supplementary Table 5.2).

Model selection for computational models of learning behavior. To capture the updating of the performance expectation ratings over time in a learning model, a similar model comparison to that of Müller-Pinzler et al. (2019) was performed. All three main models of the model space followed the idea of a Rescorla-Wagner model (Rescorla & Wagner, 1972) with one or two learning rates for each participant reflecting the degree to which people weighted prediction errors ($PE = Feedback_t - EXP_t$) to update their expectation rating (see Figure 5.3a and for model descriptions see method section).
In line with Müller-Pinzler et al. (2019), the Valence Model outperformed all other models in all three groups according to Bayesian Model Selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009; see Figure 5.3b; protected exceedance probability for the whole sample $pxp_{total} > .999$, Bayesian omnibus risk $BOR_{total} < .001$ as well as separately for the three groups $pxp_{SOC} = .985$, $BOR_{SOC} = .019$, $pxp_{PHY} > .999$, $BOR_{PHY} < .001$, $pxp_{Control} > .999$, $BOR_{Controll} < .001$; see Table 5.1 and Supplementary Table 5.3 for more details on model comparisons). This model, with two separate learning rates for positive PEs (α_{PE+}) and

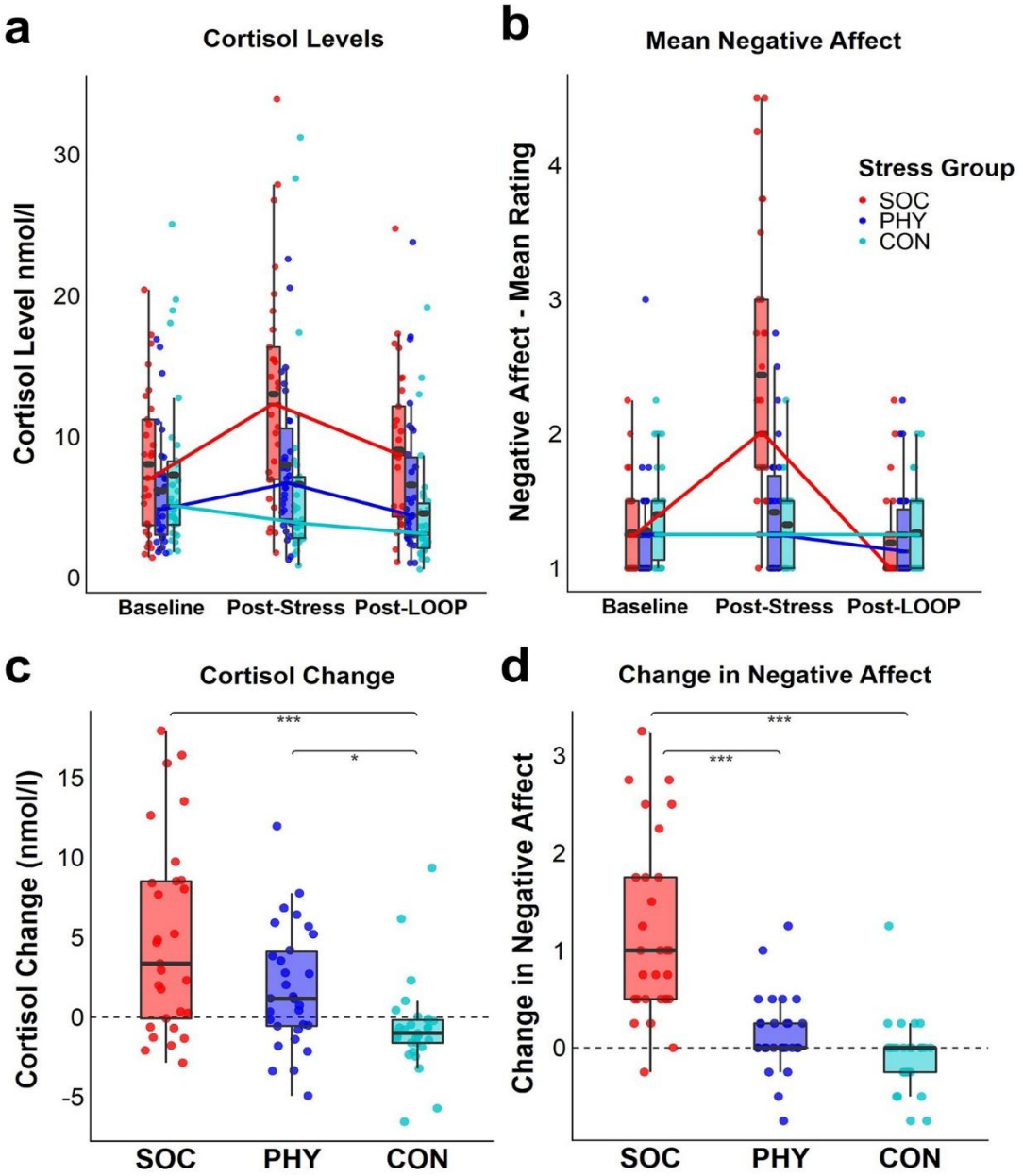


Figure 5.2. **(a)** Cortisol levels over the course of the experiment separately for the three stress groups (social-evaluative stress [*n* = 29] vs. physical stress [*n* = 30] vs. no stress [*n* = 30]). Lines connect group medians. Group means are depicted as gray ovals within the box. **(b)** Mean negative affect ratings (embarrassment, anger, frustration and perceived stress) over the course of the experiment separately for the three stress groups depicted as in **(a)**. **(c)** Change in saliva cortisol levels after stress induction (post-stress T 2_{CORT} − baseline T1_{CORT}). **(d)** Change in negative affect (post-stress T 2_{AFF} − baseline T1_{AFF}), SOC = Social-evaluative stress group, PHY = Physical stress group, CON = no stress control group. Line inside box: median, lower/upper box hinges: 25th and 75th percentile, lower/upper box whiskers: smallest/largest value within 1.5 × inter-quartile range from hinges, **p* < 0.050, ****p* < 0.001.

negative PEs (α_{PE-}) across ability conditions, assumes that learning differs depending on the valence of prediction errors. Learning parameters from the Valence Model were used for further analysis.

The modeled performance expectations of our winning model predicted the performance expectation ratings on the individual subject level within each ability condition with R^2 = 0.33 ± 0.24 (M ± SD). Repeating the model free analysis with the modeled performance

expectations confirmed the results from the original analysis (see Supplementary Table 5.4).

Table 5.1. PSIS-LOO Scores for the whole sample

Model	PSIS-LOO	LOO-SE	LOO-Diff (SE-Diff)	% of $\hat{k}$ > 0.7	No. Est. Parameters
Unity Model (M1)	-2028.5	257.0	267.1 (52.0)	0.09	3
Ability Model (M2)	-1884.4	247.4	123.0 (95.9)	0.53	4
Valence Model (M3)	-1761.4	280.4		0.17	4
Mean Model (M0)	-2531.9	219.2	770.5 (93.5)	0	2

Note. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (Valence Model) and standard errors of differences; percentage of $\hat{k}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al., 2017); No. Est. Parameters = number of estimated parameters in the model.

Stress and learning parameters. In line with Müller-Pinzler et al. (2019), the physical stress and no stress control group showed a negativity bias in their learning behavior, i.e. a stronger self-related belief updating after negative than positive prediction errors (α_{PE+} vs. α_{PE-} within group comparison for PHY: W = 100, Z = -2.73, p = .005 and CON: W = 84, Z = -2.89, p = .003, Wilcoxon test). This negativity bias was absent after social-evaluative stress (α_{PE+} vs. α_{PE-} within group comparison for SOC: W = 193, Z = -0.53, p = .609; significant PE-Valence x Contrast SOC vs. [PHY, CON] interaction $b_{VALxSOC}$ = 0.114, t_{85} = 2.30, p = .024, PE-Valence x Contrast PHY vs. CON: $b_{VALxPHY}$ = -0.036, t_{85} = -0.72, p = .471; betas standardized, see Figure 5.3d and Supplementary Table 5.5).
To better capture biased learning behavior, a valence bias score was computed (valence bias score = (α_{PE+} - α_{PE-})/(α_{PE+} + α_{PE-}); Müller-Pinzler et al., 2019; Niv, Edlund, Dayan, & O'Doherty, 2012; Palminteri, Lefebvre, Kilford, & Blakemore, 2017), which represents updating after positive compared to negative prediction errors. More positive valence bias scores indicate more self-beneficial belief updating, while negative valence bias scores speak for stronger self-related belief updating after negative feedback.

Negative affect and cortisol change predict subsequent self-beneficial belief updating. To further assess which aspect of the stress response is associated with self-beneficial belief updating, we correlated negative affect and cortisol with the valence bias scores across all three experimental groups. While both stress groups (SOC and PHY) only differ significantly in terms of an increase in negative affect but not cortisol levels, both measures show large variance within and across groups (see Figure 5.2) that could explain differences in self-related belief updating that is only partially captured by the group effect.

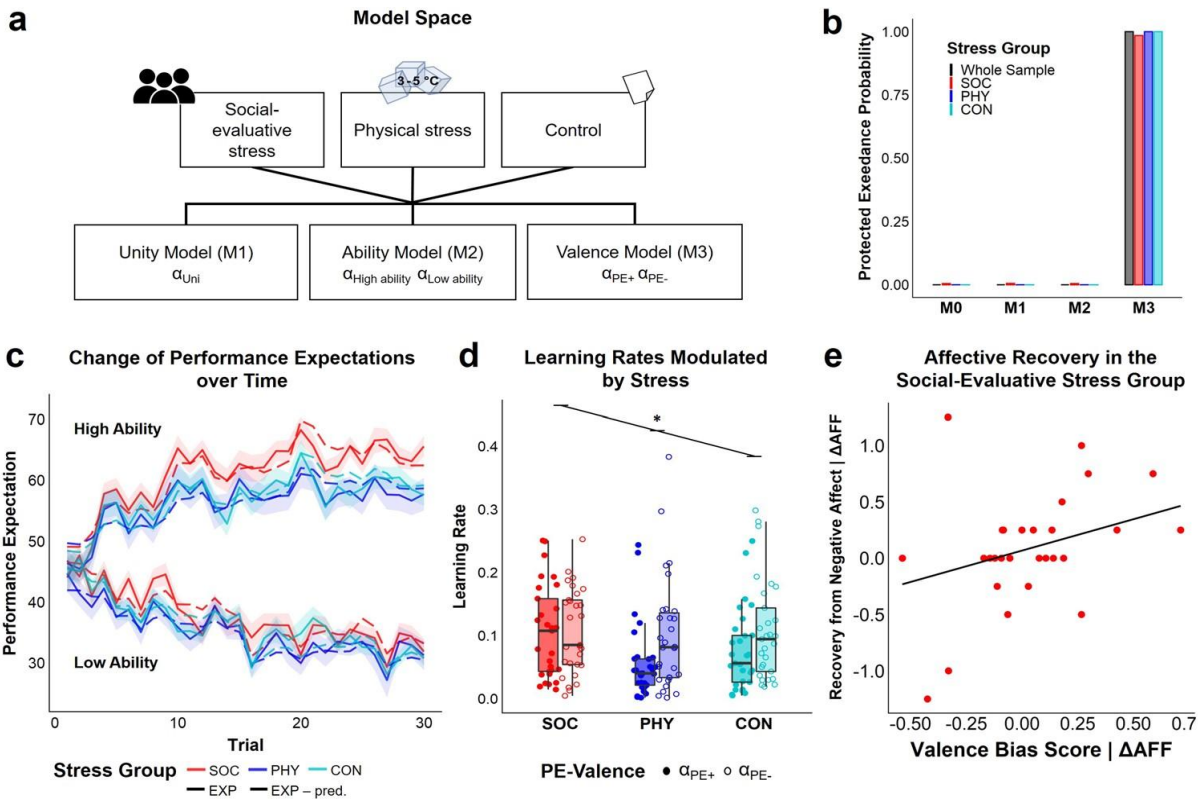


Figure 5.3. **(a)** Structure of the model space. α_{Uni} = one learning rate for the whole time course; $\alpha_{High\ ability}$/ $\alpha_{Low\ ability}$ = two separate learning rates for the two ability conditions; α_{PE+}/α_{PE-} = two separate learning rates for positive and negative prediction errors; adapted from Müller-Pinzler et al. (2019). **(b)** Protected exceedance probabilities resulting from the Bayesian Model Selection procedure including the prediction error learning models depicted in **(a)** and a mean model (M0) assuming stable means for each ability condition instead of continuous learning. **(c)** Performance expectation ratings (EXP, solid line) and performance expectations predicted by the winning model (EXP − pred., dashed line) over the time course of 30 trials. Ratings and predicted values were averaged across participants separately for the two ability conditions and the three experimental groups. Shaded areas represent the standard errors of the expectation ratings for each trial. **(d)** Learning rates derived from the Valence Model (winning model). A significant interaction effect (*) of PE-Valence x Stress group (SOC = social-evaluative stress, PHY = physical stress, CON = no stress control) indicates that a bias towards increased updating in response to negative prediction errors (α_{PE-}) in contrast to positive prediction errors (α_{PE+}) is absent in the social-evaluative stress group. **(e)** Rank-based regression plot of valence bias score predicting the recovery from negative affect (REC, ratings $T2_{AFF}$ − $T3_{AFF}$) in the subsample of the social-evaluative stress group (n = 29) controlled for the stress-induced change in negative affect (ΔAFF, ratings $T2_{AFF}$ − $T1_{AFF}$), i.e., residuals of the valence bias score and the recovery predicted by ΔAFF are plotted. More self-beneficial belief updating (higher valence bias score) is associated with a better recovery from stress-induced negative affect.

We found that a stronger increase in negative affect (ΔAFF = $T2_{AFF}$ - $T1_{AFF}$) predicted more self-beneficial belief updating ($b_{\Delta AFF}$ = 0.100, t_{86} = 2.066, p = .042, rank regression; Kloke & McKean, 2012) of the valence bias score predicted by ΔAFF for the whole sample). Also, a higher increase in cortisol levels (ΔCORT = $T2_{CORT}$ - $T1_{CORT}$) predicted more self-beneficial belief updating ($b_{\Delta CORT}$ = 0.020, t_{85} = 2.377, p = .020, rank regression; Kloke & McKean, 2012) controlled for TIME for the whole sample, for further information see 5.7 Supplementary Results and Supplementary Table 5.6).

Learning bias and affective recovery. A more positive valence bias score predicted better recovery from stress-induced negative affect during learning in the social-evaluative stress group, the only group with significantly increased levels of negative affect after stress (Figure 5.3e*;* REC, change in negative affect post-stress $T2_{AFF}$ - post-learning $T3_{AFF}$, $b_{BIAS} = 0.584$, $t_{26} = 2.131$, $p = .043$, rank regression; Kloke & McKean, 2012; controlled for the increase in negative affect [$\Delta AFF = T2_{AFF} - T1_{AFF}$]). This supports the idea of self-beneficial belief updating as a coping strategy. Analysis across the whole sample trend-wise confirmed this effect ($b_{BIAS} = 0.238$, $t_{85} = 1.922$, $p = .058$). However, regression coefficients of the valence bias score predicting the affective recovery within the two other experimental groups alone, which exhibit no substantial increase in negative affect in the first place, were not significant (PHY: $b_{BIAS} = 0.147$, $t_{27} = 0.999$, $p = 0.327$, CON: $b_{BIAS} = 0.000$, $p > 0.999$, rank regression). A stronger relationship in the social stress group compared to the other groups, i.e. a modulation of the factor Stress group, showed a trend-wise effect (BIAS x Contrast SOC vs. [PHY, CON] interaction $b_{BIASxSOC} = 0.322$, $t_{85} = 1.70$, $p = 0.093$, rank regression of the affective recovery predicted by valence bias score, Stress group split into the contrasts social [SOC] vs. non-social [PHY, CON] and PHY vs. CON, the increase in negative affect, plus the bias x Stress group interactions for the two contrasts, Supplementary Table 5.7).

5.4 Discussion

After being devalued for example at work or school we need to empower ourselves in order to uphold or boost our self-image. Research has shown that the ability to adopt a positive attitude towards oneself after receiving criticism is central to positive affect and good mental health outcomes in the long run (Leary, 2007; Roese & Olson, 2007; Taylor & Brown, 1988). In the current study we investigated how people apply self-beneficial belief updating during a performance feedback situation as a means to counter their negative affect. Using computational modelling, we provide a mechanistic explanation on how individuals engage in more self-beneficial updating of ability beliefs after experiencing a threat to their social image and how this shift in social learning of self-related information predicts recovery from stress-induced negative affect.

The positive shift of self-related updating of ability beliefs after social-evaluative stress, going along with a better recovery from negative affect, fits nicely to the notion of a belief's own value as recently posited by Bromberg-Martin and Sharot (2020). In their revised framework, general belief updating is not solely driven by external outcomes like rewards or punishments but also by the agent's motivation to optimize internal states like positive affect (Bromberg-Martin & Sharot, 2020; Stolz, Müller-Pinzler, Krach, & Paulus, 2020). In the present study we show this direct link between self-related belief updating and a change in the affective state indicating that self-related belief updating might be motivated by the wish to uphold or even recover a positive affective state. This is in line with the idea of motivated cognition, i.e. the assumption that cognitive processes like attention, information processing and decision making are not neutral on their own, but are always shaped by needs, feelings and desires of the individual (Hughes & Zaki, 2015). Especially when processing information that challenges one's self-image, self-related belief updating

is not only informed by the history of previous feedback, as it has often been assumed in classical reinforcement learning tasks, but also by various self-relevant needs and goals (Kuzmanovic & Rigoux, 2017). Transferred to the present study, this implies that the motivation to restore an endangered self-image and to regulate one's affect back to a set point directly impacts self-related information processing. The pattern of an active counter-regulation of negative affect by self-beneficial belief updating can be described as a striving for homeostasis (Roese & Olson, 2007). To better capture the fluctuation of the affective state and its involvement in the trial-by-trial self-related belief updating loop, following the framework by Bromberg-Martin and Sharot (Bromberg-Martin & Sharot, 2020), future studies should consider repeated assessments of affective states during the task to predict the empowering potential of shifts in learning on the single trial level.

Since negative self-related beliefs are at the core of psychiatric conditions like depression (Beck, 2002), this study targets clinically highly relevant processes. Depression is associated with seeking negative feedback which confirms negative self-related beliefs (Giesler, Josephs, & Swann, 1996) and seeking negative feedback in combination with a stressful life event can further increase depressive symptoms (Pettit & Joiner, 2001). Furthermore, depression is associated with a weaker stress recovery mediated by an attentional bias towards negative feedback. Understanding the mechanisms of how people form self-related beliefs in a context mimicking everyday performance settings and linking these to the regulation of negative affect after stress has important implications for understanding the etiology of depressive symptoms. The present study set-up, including a social-evaluative stress induction followed by a social-evaluative performance situation also addresses one of the fundamental fears of individuals with social anxiety: being devalued by others. Since both depression and social anxiety are associated with negatively biased updating behavior in response to self-related feedback (Gotlib & Krasnoperova, 1998; Koban et al., 2017; Korn et al., 2014; Müller-Pinzler et al., 2019), we assume that the affect-regulating and empowering potential of self-beneficial belief updating after social-evaluative stress would be less pronounced in depression or social anxiety and would thereby possibly exacerbate the symptomatology in a self-fulfilling way. Future studies with similar experimental set-ups and clinical samples could examine the relationship between self-beneficial belief updating and affect regulation in more detail and develop potential intervention strategies based on empowering individuals on their way to processing newly incoming information.

Replicating a previous study of ours (Müller-Pinzler et al., 2019), self-related belief updating was negatively biased in the control condition in which participants were not exposed to any stress. In the prior study, this negativity bias has been shown to be specific for self-related belief updating in comparison to belief updating about another person (Müller-Pinzler et al., 2019). In the present study, we found that after physical stress participants also exhibited a negativity bias in forming self-related beliefs, i.e. participants tended to make greater updates in response to negative prediction errors in contrast to positive prediction errors. The negativity bias stands in contrast to other studies reporting a positivity or optimism bias in feedback-based learning e.g. when receiving feedback

about the chance to encounter negative life events (Kuzmanovic, Jefferson, & Vogeley, 2016; Sharot et al., 2011), about one's intelligence or about one's personality (Eil & Rao, 2010; Korn et al., 2012; for a review see Sharot & Garrett, 2016). There are several possible explanations for the motivation behind the negativity bias in context of the LOOP task in contrast to the reported positivity biases of other studies which was, however, not the focus of the present study (for a discussion on the negativity bias see Müller-Pinzler et al., 2019). In order to test for the specificity of self-beneficial belief updating after social-evaluative stress, it would be interesting to test if this effect also accounts for experiments that typically yield a positivity bias (e.g. for life events, IQ or personality) in feedback-based learning tasks.

Here, we demonstrated that both, negative affect and cortisol stress responses, go along with a shift in self-related belief updating. It has been shown before that experiencing social emotions (e.g. embarrassment or shame) is related to increased cortisol levels in situations which threaten one's social image, like the social-evaluative stress induction (Gruenewald et al., 2004). Cortisol has been linked to reward processing and feedback-based learning in the *stress triggers additional reward salience - STARS* - model which proposes that stress and the associated release of cortisol modulates the dopamine system, resulting in an increased salience of rewards, thus biasing learning towards rewarding feedback (Lighthall et al., 2013; Mather & Lighthall, 2012). The current results, however, suggest that the quality of stress (here, social vs. physical) might make a difference, and the STARS model, based on a rather unspecifically triggered cortisol response, cannot fully explain the present stress effect on self-related belief updating after social but not physical stress.

Although both groups showed a significantly greater increase in cortisol compared to control, this effect was less pronounced in the physical stress group. While the Cold Pressor Test is known to elicit a strong sympathetic activity, studies reported only low to moderate cortisol effects (Schwabe, Haddad, & Schachinger, 2008), that were weaker than after the Trier Social Stress Test (McRae et al., 2006; Skoluda et al., 2015). Our alteration of the Cold Pressor Test, to remove the social element of the conductor, might have even further reduced stress effects as compared to the original Cold Pressor Test protocol. We did find an association of the negative affect response (i.e. self-evaluative emotions like embarrassment), which is typically specific for social-evaluative stress and rather absent during physical stress, with shifts in learning behavior in our study. But also cortisol as a rather unspecific stress component was associated with shifts in learning behavior. Thus, we cannot rule out that a more intense physical stress protocol with higher cortisol responses would have led to similar effects on learning rates. A more detailed recording of negative affect and a comparison between different negative affective states as well as more detailed recording of the physiological stress response might help in future studies to better differentiate between different stress qualities and understand specific effects of social-evaluative stress on self-beneficial belief updating.

To summarize, our results indicate a shift towards more self-beneficial belief updating after social-evaluative but not physical stress. This shift goes along with a better recovery from

stress-induced negative affect. Linking self-related belief updating to affect is an important step in understanding biases in self-related learning and its relation to affect regulation. The special feature of the present study was the study-set that allowed to examine a link between negative affect and self-related belief updating. By introducing a performance context with consecutive self-related feedback, corresponding to real-life school or work related performance situations, individuals can form beliefs about their own abilities over time and potentially use this formation process as a means to regulate their affect. With this approach we aimed to increase the ecological validity of the study in order to trigger and investigate motivational processes that might be less relevant in more abstract study settings. Since social evaluation represents a constant stressor in every-day life, the question of an appropriate coping strategy to regulate negative affect is of great importance when handling everyday social situations.

5.5 Materials and Methods

Participants. Eighty-nine participants recruited at the University of Lübeck Campus were included in the study. Upon appearance, participants were assigned to either a social-evaluative stress group (SOC; n = 29, 21 female, aged 18–28 years; M = 22.9; SD = 2.76), a physical stress group (PHY; n = 30, 20 female, aged 19–27 years; M = 22.5; SD = 1.94) or the control group (CON; n = 30, 20 female, aged 18–32 years; M = 22.3; SD = 3.00, data of the control group were published before, Müller-Pinzler et al., 2019). From the initially recruited N = 96 subjects, seven had to be excluded – five because they did not believe the cover story and two due to technical problems. All included participants were fluent in German, non-smokers with a body-mass index between 18.5 and 30. They were not diagnosed with acute or chronic psychiatric conditions or diseases affecting the hormone system and did not take psychiatric drugs or medication affecting the hormone system (except hormonal contraceptives). Participants had normal or corrected-to-normal vision and did not study psychology to avoid previous experience with experiments using cover stories. Additional exclusion criteria for participants who underwent the physical stress protocol were cardiovascular diseases, frequent fainting or seizures and current hand injuries. For more details on the sample characteristics see Supplementary Table 5.9a. All participants gave written informed consent prior to the participation and received monetary compensation for their participation. They were naive to the background of the study during the session and debriefed about the cover story afterwards. The study was conducted in compliance with the ethical guidelines of the American Psychological Association (APA) and was approved by the ethics committee of the University of Lübeck.

Manipulation procedure. *Social-evaluative stress.* Social-evaluative stress was induced by a public speech similarly to the Trier Social Stress Test (Kirschbaum et al., 1993). Participants were instructed to prepare a short self-presentation for an application for a scholarship, which had to be presented in front of a selection committee who would allegedly assess the participant's verbal skills and body language. The selection committee consisted of the experimenter, who was passive during the speech, a second

experimenter, who was allegedly responsible for measuring verbal skills, and a passive camera assistant, who pretended to videotape the speech. Before starting the 10-min preparation period, participants briefly visited the room with the selection committee. After the preparation time was over, participants were asked to come back to this room and present their speech. Talking time was 5 min (*M* = 4.9 min, *SD* = 0.16) with a minimum of 3 min of uninterrupted speech. If the participant finished the speech before the time was over, the second experimenter waited for at least 15 s with a motionless face and then asked the participant to continue. If the participant stopped speaking again and the 3 min of free speech had passed, the second experimenter asked standardized questions until the 5 min of talking time were over ("Explain why it is important for you to achieve a good performance.", "Do you think it is important to improve yourself throughout your life?", "Do you consider yourself a person who values his/her independence?"). Average social-evaluative stress duration (start subsequent rest period – start speech preparation) was *M* = 16.4 min, *SD* = 1.2.

Physical stress. Physical stress was induced by an exposure to ice water according to the Cold Pressor Test protocol (Hines & Brown, 1932, 1933). Participants were asked to dip their non-dominant hand in cold water (water temperature 3 – 5.5°C = 37.4 – 41.9°F, *M* = 4.26°C, *SD* = 0.50) for as long as possible up to 3 min (duration 48 s – 3 min, *M* = 2.7 min, *SD* = 0.7). The water was kept in motion with a small electrical pump to prevent the water temperature from rising around the participant's hand. To control for the procedure of the social-evaluative stress condition, participants visited the room with the cold pressor apparatus first, had a 10-min preparation period and came back into the room for the stress exposure. During the preparation time, participants were asked to imagine dipping their hands in a freezing cold environment and write down their associations. To make the stress exposure less social, the experimenter was not present in the room but waited in an adjacent room. If the participant took out their hand before the three minutes were over, they had to signal this immediately by ringing a bell. The experimenter could roughly observe the participant in the reflection of the glass door, thus ensuring that she/he dipped the hand into the water. Average physical stress duration (including preparation period) was *M* = 16.2 min, *SD* = 1.5.

No stress control condition. In the control condition, participants performed a reading task that was described to them as measuring reading speed. They had 10 min to rehearse two different texts about applying for a scholarship. Afterwards, they were guided to the other room with nobody present and were asked to measure their reading time, while reading the two texts aloud at a natural speed. Average control duration was *M* = 15.3 min, *SD* = 1.3.

Manipulation checks. *Cortisol.* Three saliva samples were collected during the experiment for cortisol analysis (see Figure 5.1a). The first sample (baseline $T1_{CORT}$) was taken after a 10 min period of rest immediately before starting the instruction for the stress manipulation (mean time between $T1_{CORT}$ and start of the SOC, PHY or CON preparation phase: *M* = 3.7 min, *SD* = 1.4). The post-stress cortisol sample $T2_{CORT}$ was collected after another 10

min resting period following the stress manipulation and the last sample ($T3_{CORT}$) was collected after the learning task (*M* = 45.6 min (*SD* = 3.3) post stress). The stress-induced cortisol change (ΔCORT) was determined by subtracting the cortisol levels of $T2_{CORT}$ - $T1_{CORT}$. Saliva was collected with Salivettes (Sarstedt, Nümbrecht, Germany), stored at -30 °C and sent to the bio-psychological lab at TU Dresden, Dresden, Germany for analysis (here stored at -20 °C until analysis). Salivary free cortisol levels were determined using a chemoluminescence immunoassay (IBL International, Hamburg, Germany).

Negative affect. We assessed negative affect by means of a short pen and paper questionnaire, covering the emotions embarrassment, anger, frustration, as well as the perceived stress with one rating each. The questionnaires were handed out at baseline ($T1_{AFF}$) as well as at the very end of the experiment ($T3_{AFF}$). The post-stress negative affect was measured immediately after the stress manipulation ($T2_{AFF}$; see Figure 5.1). Ratings were averaged for each measurement point to get a composite measure of negative affect (see Supplementary Figure 5.1 for separate scores). The change in negative affect after stress (ΔAFF) was determined by subtracting T1 negative affect from T2 ($T2_{AFF}$ - $T1_{AFF}$). The recovery from negative affect (REC) was determined by subtracting T3 negative affect from T2 ($T2_{AFF}$ - $T3_{AFF}$).

Behavioral task. *Learning of own performance task.* The Learning of own performance (LOOP) task (Müller-Pinzler et al., 2019; Figure 5.1b) allows to measure self-related belief updating through trial-by-trial performance expectation ratings and subsequent performance feedback. The task included estimation questions in two different estimation categories (heights of houses and weights of animals) and was presented to the participants as a measure of estimation abilities. To make participants learn about their estimation ability the two estimation categories were paired with manipulated performance feedback implying high ability for one category and low ability for the other (e.g. heights of houses = high ability and weights of animals = low ability, estimation categories were counterbalanced between ability conditions). The assignment of the categories to the ability conditions was independent of the participants' actual performance and their performance expectation ratings. Thus, participants could learn over the course of the experiment that they were good in one estimation category and rather bad in the other one. Each trial began with a cue displaying the category of the next estimation question followed by a performance expectation rating for this question. Afterwards, the estimation question was presented together with a picture for 10 s. Continuous response scales below the pictures determined a range of plausible answers for each question, and participants indicated their responses by navigating a pointer on the response scale with a computer mouse. Subsequently, feedback indicating the estimation accuracy as percentiles compared to an alleged reference group of 350 university students was presented for 5 s (e.g. "You are better than 72 % of the reference participants."). The order of the two estimation categories/ability conditions was intermixed with a maximum of two consecutive trials of the same condition and 30 trials per condition in total. The estimation questions were randomized within the estimation category/ability conditions. A

fixed sequence of ability conditions and feedback was presented for all participants. In the low ability condition, feedback was approximately normally distributed around the 35th percentile (SD ≈ 16; range 1–60%) and in the high ability condition around the 65th percentile (SD ≈ 16; range 40–99%). The task started with detailed instructions and three test trials. All stimuli were presented using MATLAB Release 2015b (The MathWorks, Inc.) and the Psychophysics Toolbox (Brainard, 1997).

Procedure. To minimize noise in the cortisol saliva samples, participants were asked to follow behavioral rules prior to the experimental session. These were in detail: no alcohol on the evening before the experiment and bed rest at about 10 p.m. (ideal case eight hours of sleep); one hour before the session: no sport, no smoking, no drinks containing caffeine or theine, no food (including bonbons and chewing gums) and no juices. Upon arrival at the laboratory, participants read the participant information including the cover story regarding the stress manipulation and the LOOP task. After signing the consent form, they were asked to fill out a questionnaire checking the adherence to the behavioral rules. Participants rested for 10 min before the baseline measurement, including saliva cortisol and negative affect, was obtained ($T1_{AFF/CORT}$). During the resting period, they filled out a short personality questionnaire (not included in this study). Subsequently, participants of the social and physical stress groups were challenged with a stress protocol while participants of the control group did the control reading task. Directly afterwards, participants rated their affective state ($T2_{AFF}$) followed by another 10 min resting period, which was terminated with a saliva sampling ($T2_{CORT}$). In the second part of the experiment participants performed the LOOP task. Finally, another cortisol sample and affective ratings were collected ($T3_{AFF/CORT}$). After completing a post-experimental interview, including additional questionnaires, participants were debriefed about the cover story. The experimental sessions were run between 10.00 a.m. - 12.00 p.m., 1.00 - 3.00 p.m. or 3.45 - 5.45 p.m. The allocation to the time slots did not differ between the experimental groups (Pearson's Chi-squared test p = .867, see Supplementary Table 5.10b). See Figure 5.1a for a graphical illustration of the procedure.

Statistical analysis. *Stress manipulation.* To test whether the stress manipulation was effective, the stress-induced changes in cortisol as well as affect were compared between the three experimental groups. Due to the stress manipulation, the variance of the cortisol and negative affect responses were unequal between the three experimental groups (Levene test *p*s < .05). Since the distributions of the cortisol and affective stress response were skewed in some groups (Lilliefors-corrected Kolmogorov-Smirnov normality test *p*s < .05 for the cortisol change in the control group and for the change in negative affect in all groups) non-parametric tests were used. Since cortisol levels are known to underlie circadian fluctuations (Weitzman et al., 1971) all cortisol analysis were controlled for time of the day (morning vs. noon vs. afternoon, see Procedure). Responses in negative affect were compared with the Kruskal-Wallis test, the cortisol response was compared with the Scheirer-Ray-Hare test, an extension of the Kruskal–Wallis test that allows to control for

time of the day. Post-hoc comparisons between the groups were performed with Dunn's test.

Model free analysis of performance expectation ratings. The analysis of the expectation ratings including computational modeling was adapted from Müller-Pinzler et al. (2019). To illustrate basic effects of the expectation ratings, a linear mixed model with the factors Ability condition (high ability vs. low ability), the continuous variable Trial (30 Trials), and Stress group (with the two contrasts SOC vs. [PHY, CON] and PHY vs. CON) as a between subject factor was performed.

Computational modeling of learning behavior. The dynamic changes in self-related beliefs, which were measured by the performance expectation ratings in response to the provided performance feedback, were modeled using prediction error delta-rule update equations (adapted from Rescorla-Wagner model; Rescorla & Wagner, 1972). There were three main models of the model space with one or two learning rates modeled separately for each participant (see Figure 5.3a). The first model (Unity Model) included a single learning rate for the whole time course ($EXP_{t+1} = EXP_t + \alpha_{Uni}\ PE_t$). The second model (Ability Model) contained two separate learning rates for the two ability conditions allowing to capture a difference in expectation updating when receiving feedback in a high ability context ($\alpha_{HighAbility}$) or low ability context ($\alpha_{LowAbility}$). The third model (Valence Model) with two separate learning rates for positive PEs (α_{PE+}) and negative PEs (α_{PE-}) across ability conditions allows to model learning that differs depending on the valence of prediction errors rather than different ability conditions. The three models were compared to a Mean Model with two performance expectations means reflecting the assumption of stable expectations for each ability condition without learning over time. In addition to the learning rates, we fitted two parameters for the initial belief about the participant's performance, separately for both ability conditions (see Table 5.1).

Model fitting. For model fitting we used the RStan package (Stan Development Team, 2019), which uses Markov chain Monte Carlo (MCMC) sampling algorithms. All learning models of the model space were fitted separately for each subject. To sample posterior parameter distributions, a total of 2400 samples were drawn after 1000 burn-in samples (overall 3400 samples; thinned with a factor of 3) in three MCMC chains. Convergence of the MCMC chains to the target distributions was assessed by $\hat{R}$ values (Gelman & Rubin, 1992) for all model parameters. One subject was excluded due to implausible model parameters, i.e. mean learning rate of almost 1, as well as $\hat{R}$ values of 1.1 and low effective sample sizes (n_{eff}, estimates of the effective number of independent draws from the posterior distribution) for some model parameters of the valence model. Otherwise the effective sample sizes were greater than 1000 (>1400 for most parameters). Posterior distributions for all parameters for each of the participants were summarized by their mean resulting in a single parameter value per subject that we used to calculate group statistics.

Bayesian model selection and family inference. To select the model that describes the participants' updating behavior best, we estimated pointwise out-of-sample prediction accuracy for all fitted models separately for each participant by approximating leave-one-

out cross-validation (LOO; Vehtari, Gelman, & Gabry, 2017). To this end, we applied Pareto-smoothed importance sampling (PSIS) using the log-likelihood calculated from the posterior simulations of the parameter values as implemented by Vehtari et al. (2017; loo *R* package; Vehtari, Gabry, Magnusson, Yao, & Gelman, 2019). Sum PSIS-LOO scores for each model as well as information about $\hat{k}$ values, the estimated shape parameters of the generalized Pareto distribution, indicating the reliability of the PSIS-LOO estimate, are depicted in Table 5.1. As summarized in Table 5.1, very few trials resulted in insufficient parameter values for $\hat{k}$ and thus potentially unreliable PSIS-LOO scores (on average 0.20 % of trials per subject with $\hat{k} > 0.7$). Bayesian model selection on PSIS-LOO scores was performed on the group level accounting for group heterogeneity as described by Stephan et al. (2009). This procedure provides the protected exceedance probability for each model (*pxp*), indicating how likely a given model has a higher probability explaining the data than all other models, as well as the Bayesian omnibus risk (*BOR*), the posterior probability that model frequencies for all models are all equal to each other (Rigoux et al., 2014). Additionally, difference scores of PSIS-LOO for all models in contrast to the winning model were computed, which can be interpreted as a simple 'fixed-effect' model comparison (Vehtari et al., 2017; see Table 5.1).

Posterior predictive checks. To test whether the predicted values of the winning model could capture the variance in the performance expectation ratings a regression analysis (EXP ~ pred. values) was performed for each subject separately for the two ability conditions. R-squared statistic was determined and averaged. In addition, the model free analysis of the expectation ratings was repeated with the predicted values of the winning model to assess if the predicted data captured the effects that were present in the data of the expectation ratings.

Analysis of learning parameters. Learning rates for positive (α_{PE+}) and negative prediction errors (α_{PE-}, factor PE-Valence) were compared between the three groups in a linear mixed model with the factors PE-Valence and group (split into the contrasts SOC vs. [PHY, CON] and PHY vs. CON). Additional post-hoc tests for the PE-Valence within each stress group were performed with the Wilcoxon test. To test whether the variance in affective response and the cortisol response created by our stress manipulation is related to a bias in the updating behavior, we calculated a normalized learning rate valence bias score (valence bias score = $(\alpha_{PE+} - \alpha_{PE-})/ (\alpha_{PE+} + \alpha_{PE-})$; Müller-Pinzler et al. 2019) and tested whether the affective response and the cortisol response predicted this bias. Since the distribution of the cortisol and affective response is left-skewed due to the experimental manipulation and thus absence of the response in parts of the subjects (valence bias score is normally distributed in the whole sample as well as all subgroups) we used rank regressions following Kloke et al. (2012). In case of the cortisol response, time of the day was additionally included in the regression analysis as a control variable to take into account circadian fluctuations of cortisol levels. To test whether the learning bias is associated with the recovery from negative affect elicited by stress (change in affective ratings post-stress $T2_{AFF}$ – post-learning $T3_{AFF}$), rank regressions with the valence bias score predicting the

recovery were computed with the stress-induced increase in negative affect as an additional control variable to take into account regression to the mean. This was computed within the three experimental groups as well as for the whole sample, here with the additional variable group (split into the contrasts SOC vs. [PHY, CON] and PHY vs. CON) as well as the interaction bias x group to test whether the correlation differs between the three experimental groups. Data was analyzed in with the software R version 3.6.0 (R Core Team, 2013) and plots were made with the R package gglpot2 (Wickham, 2016).

5.6 References

Adler, C. M., Elman, I., Weisenfeld, N., Kestler, L., Pickar, D., & Breier, A. (2000). Effects of acute metabolic stress on striatal dopamine release in healthy volunteers. *Neuropsychopharmacology, 22*(5), 545–550. https://doi.org/10.1016/S0893-133X(99)00153-0

Adler, C. M., Elman, I., Weisenfeld, N., Kestler, L., Pickar, D., & Breier, A. (2000). Effects of acute metabolic stress on striatal dopamine release in healthy volunteers. *Neuropsychopharmacology*, *22*(5), 545–550. https://doi.org/10.1016/S0893-133X(99)00153-0

Antov, M. I., Wölk, C., & Stockhorst, U. (2013). Differential impact of the first and second wave of a stress response on subsequent fear conditioning in healthy men. *Biological Psychology*, *94*(2), 456–468. https://doi.org/10.1016/j.biopsycho.2013.08.007

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Beck, A. T. (2002). Cognitive models of depression. In R. L. Leahy & E. T. Dowd (Eds.), *Clinical Advances in Cognitive Psychotherapy: Theory and Application* (14 (1), pp. 29–61). New York: Springer Publishing Company.

Bogdan, R., & Pizzagalli, D. A. (2006). Acute stress reduces reward responsiveness: implications for depression. *Biological Psychiatry*, *60*(10), 1147–1154. https://doi.org/10.1016/j.biopsych.2006.03.037

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Bromberg-Martin, E. S., & Sharot, T. (2020). The Value of Beliefs. *Neuron*, *106*(4), 561–565. https://doi.org/10.1016/j.neuron.2020.05.001

Burke, P. J. (1991). Identity Processes and Social Stress. *American Sociological Review*, *56*, 836–849.

Campbell, J., & Ehlert, U. (2012). Acute psychosocial stress: Does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology*, *37*(8), 1111–1134. https://doi.org/10.1016/j.psyneuen.2011.12.010

Eil, D., & Rao, J. M. (2010). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138.

Eisenberger, N. I., Inagaki, T. K., Muscatell, K. A., Haltom, K. E. B., & Leary, M. R. (2011). The neural sociometer: Brain mechanisms underlying state self-esteem. *Journal of Cognitive Neuroscience*, *23*(11), 3448–3455. https://doi.org/10.1162/jocn_a_00027

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior and Organization*, *80*(3), 532–545. https://doi.org/10.1016/j.jebo.2011.05.012

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. *The Journal of Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.1523/JNEUROSCI.0716-18.2018

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Giesler, R. B., Josephs, R. A., & Swann, W. B. (1996). Self-verification in clinical depression: The desire for negative evaluation. *Journal of Abnormal Psychology*, *105*(3), 358–368. https://doi.org/10.1037/0021-843X.105.3.358

Glanz, K., & Schwartz, M. D. (2008). Stress, Coping, and Health Behavior. In *Health Behavior and Health Education* (4th ed., pp. 211–236). San Francisco: Jossey-Bass.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(SUPPL. 3), 15647–15654. https://doi.org/10.1073/pnas.1014269108

Gloria, C. T., & Steinhardt, M. A. (2016). Relationships Among Positive Emotions, Coping, Resilience and Mental Health. *Stress and Health*, *32*(2), 145–156. https://doi.org/10.1002/smi.2589

Gotlib, I. H., & Krasnoperova, E. (1998). Biased information processing as a vulnerability factor for depression. *Behavior Therapy*, *29*(4), 603–617. https://doi.org/10.1016/S0005-7894(98)80020-8

Greenberg, J., Pyszczynski, T., Burling, J., & Tibbs, K. (1992). Depression, self-focused attention, and the self-serving attributional bias. *Personality and Individual Differences*, *13*(9), 959–965. https://doi.org/10.1016/0191-8869(92)90129-D

Gruenewald, T. L., Kemeny, M. E., Aziz, N., & Fahey, J. L. (2004). Acute threat to the social self: shame, social self-esteem, and cortisol activity. *Psychosomatic Medicine*, *66*(6), 915–924. https://doi.org/10.1097/01.psy.0000143639.61693.ef

Hines, E. A., & Brown, G. E. (1932). A standard stimulus for measuring vasomotor reactions: its application in study of hypertension. *Proceedings of the Staff Meetings of the Mayo Clinic*, *7*, 332–335.

Hines, E. A., & Brown, G. E. (1933). A standard test for measuring the variability of blood pressure: its significance as an index of the prehypertensive state. *Annals of Internal Medicine*, *7*(2), 209–217. https://doi.org/10.7326/0003-4819-7-2-209

Holly, E. N., & Miczek, K. A. (2016, January 1). Ventral tegmental area dopamine revisited: Effects of acute and repeated stress. *Psychopharmacology*, Vol. 233, pp. 163–186. https://doi.org/10.1007/s00213-015-4151-3

Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19*(2), 62–64. https://doi.org/10.1016/j.tics.2014.12.006

Joëls, M., & Baram, T. Z. (2009, June 2). The neuro-symphony of stress. *Nature Reviews Neuroscience*, Vol. 10, pp. 459–466. https://doi.org/10.1038/nrn2632

Joëls, M., Pu, Z., Wiegert, O., Oitzl, M. S., & Krugers, H. J. (2006). Learning under stress: how does it work? *Trends in Cognitive Sciences*, *10*(4), 152–158. https://doi.org/10.1016/j.tics.2006.02.002

Jundt, D. K., & Hinsz, V. B. (2002). Influences of positive and negative affect on decisions involving judgmental biases. *Social Behavior and Personality*, *30*(1), 45–52. https://doi.org/10.2224/sbp.2002.30.1.45

Kessler, R. C., Price, R. H., & Wortman, C. B. (1985). Social factors in psychopathology: Stress, social support, and coping pro-cesses. *Annu. Rev. Psychol.*

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' – A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1–2), 76–81. https://doi.org/10.1159/000119004

Kloke, J. D., & McKean, J. W. (2012). Rfit: Rank-based estimation for linear models. *R Journal*, *4*(2), 57–64. https://doi.org/10.32614/rj-2012-014

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., … Arch, J. J. (2017). Social Anxiety is Characterized by Biased Learning About Performance and the Self. *Emotion*. https://doi.org/10.1037/emo0000296

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592. https://doi.org/10.1017/S0033291713001074

Kumar, P., Berghorst, L. H., Nickerson, L. D., Dutra, S. J., Goer, F. K., Greve, D. N., & Pizzagalli, D. A. (2014). Differential effects of acute stress on anticipatory and consummatory phases of reward processing. *Neuroscience*, *266*, 1–12. https://doi.org/10.1016/j.neuroscience.2014.01.058

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162. https://doi.org/10.1016/j.neuroimage.2016.02.014

Kuzmanovic, B., & Rigoux, L. (2017). Valence-dependent belief updating: Computational validation. *Frontiers in Psychologysychology*, *8*(Jun 29), 1087. https://doi.org/10.3389/fpsyg.2017.01087

Lazarus, R. S., & Folkman, S. (1984). *Stress, Appraisal, and Coping*. New York: Springer Publishing Company.

Leary, M. R. (2007). Motivational and Emotional Aspects of the Self. *Annual Review of Psychology*, *58*(1), 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Lighthall, N. R., Gorlick, M. A., Schoeke, A., Frank, M. J., & Mather, M. (2013). Stress modulates reinforcement learning in younger and older adults. *Psychology and Aging*, *28*(1), 35–46. https://doi.org/10.1037/a0029823

Luethi, M., Meier, B., & Sandi, C. (2009). Stress effects on working memory, explicit memory, and implicit memory for neutral and emotional stimuli in healthy men. *Frontiers in Behavioral Neuroscience*, *3*(JAN), 1–9. https://doi.org/10.3389/neuro.08.005.2008

Markus, H. R., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology*, *38*(1), 299–337. https://doi.org/10.1146/annurev.psych.38.1.299

Mather, M., & Lighthall, N. R. (2012). Both risk and reward are processed differently in decisions made under stress. *Current Directions in Psychological Science*, *21*(2), 36–41. https://doi.org/10.1177/0963721411429452.

McRae, A. L., Saladin, M. E., Brady, K. T., Upadhyaya, H., Back, S. E., & Timmerman, M. A. (2006). Stress reactivity: biological and subjective responses to the cold pressor and Trier Social stressors. *Human Psychopharmacology: Clinical and Experimental*, *21*(6), 377–385. https://doi.org/10.1002/hup.778

Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing Self-Confidence: Theory and Experimental Evidence. *NBER Working Papers*. Retrieved from https://ideas.repec.org/p/nbr/nberwo/17014.html

Müller-Pinzler, L., Czekalla, N., Mayer, A. V., Stolz, D. S., Gazzola, V., Keysers, C., … Krach, S. (2019). Negativity-bias in forming beliefs about own abilities. *Scientific Reports*, *9*(1), 14416. https://doi.org/10.1038/s41598-019-50821-w

Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., … Krach, S. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, *119*, 252–261. https://doi.org/10.1016/j.neuroimage.2015.06.036

Nikolova, Y. S., Bogdan, R., Brigidi, B. D., & Hariri, A. R. (2012). Ventral striatum reactivity to reward and recent life stress interact to predict positive affect. *Biological Psychiatry*, *72*(2), 157–163. https://doi.org/10.1016/j.biopsych.2012.03.014

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/jneurosci.5498-10.2012

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*(8), e1005684. https://doi.org/10.1371/journal.pcbi.1005684

Payer, D., Williams, B., Mansouri, E., Stevanovski, S., Nakajima, S., Le Foll, B., … Boileau, I. (2017). Corticotropin-releasing hormone and dopamine release in healthy individuals. *Psychoneuroendocrinology*, *76*, 192–196. https://doi.org/10.1016/j.psyneuen.2016.11.034

Pettit, J., & Joiner, T. E. (2001). Negative-Feedback Seeking Leads to Depressive Symptom Increases under Conditions of Stress. *Journal of Psychopathology and Behavioral Assessment*, *23*(1), 69–74. https://doi.org/10.1023/A:1011047708787

Petzold, A., Plessow, F., Goschke, T., & Kirschbaum, C. (2010). Stress reduces use of negative feedback in a feedback-based learning task. *Behavioral Neuroscience*, *124*(2), 248–255. https://doi.org/10.1037/a0018930

Porcelli, A. J., Lewis, A. H., & Delgado, M. R. (2012). Acute stress influences neural circuits of reward processing. *Frontiers in Neuroscience*, *6*(NOV), 1–9. https://doi.org/10.3389/fnins.2012.00157

Qin, S., Hermans, E. J., van Marle, H. J. F., Luo, J., & Fernández, G. (2009). Acute Psychological Stress Reduces Working Memory-Related Activity in the Dorsolateral Prefrontal Cortex. *Biological Psychiatry*, *66*(1), 25–32. https://doi.org/10.1016/j.biopsych.2009.03.006

R Core Team. (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Vienna, Austria.

Raio, C. M., Hartley, C. A., Orederu, T. A., Li, J., & Phelps, E. A. (2017). Stress attenuates the flexible updating of aversive value. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), 11241–11246. https://doi.org/10.1073/pnas.1702565114

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement. In A. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Robinson, O. J., Overstreet, C., Charney, D. R., Vytal, K., & Grillon, C. (2013). Stress increases aversive prediction error signal in the ventral striatum. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(10), 4129–4133. https://doi.org/10.1073/pnas.1213923110

Roese, N. J., & Olson, J. M. (2007). Better, Stronger, Faster: Self-Serving Judgment, Affect Regulation, and the Optimal Vigilance Hypothesis. *Perspectives on Psychological Science*, *2*(2), 124–141. https://doi.org/10.1111/j.1745-6916.2007.00033.x

Rohleder, N., Beulen, S. E., Chen, E., Wolf, J. M., & Kirschbaum, C. (2007). Stress on the dance floor: the cortisol stress response to social-evaluative threat in competitive ballroom dancers. *Personality and Social Psychology Bulletin*, *33*(1), 69–84. https://doi.org/10.1177/0146167206293986

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schwabe, L., Haddad, L., & Schachinger, H. (2008). HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*, *33*(6), 890–895. https://doi.org/10.1016/j.psyneuen.2008.03.001

Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience and Biobehavioral Reviews*, *36*(7), 1740–1749. https://doi.org/10.1016/j.neubiorev.2011.07.002

Schwabe, L., & Wolf, O. T. (2012). Stress modulates the engagement of multiple memory systems in classification learning. *Journal of Neuroscience*, *32*(32), 11042–11049. https://doi.org/10.1523/JNEUROSCI.1484-12.2012

Sharot, T. (2011, December 6). The optimism bias. *Current Biology*, Vol. 21, pp. R941–R945. https://doi.org/10.1016/j.cub.2011.10.030

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33. https://doi.org/10.1016/j.tics.2015.11.002

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. https://doi.org/10.1038/nn.2949

Skoluda, N., Strahler, J., Schlotz, W., Niederberger, L., Marques, S., Fischer, S., … Nater, U. M. (2015). Intra-individual psychological and physiological responses to acute laboratory stressors of different intensity. *Psychoneuroendocrinology*, *51*, 227–236. https://doi.org/10.1016/j.psyneuen.2014.10.002

Stan Development Team. (2019). *RStan: the R interface to Stan, R package version 2.19.2*.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Stolz, D. S., Müller-Pinzler, L., Krach, S., & Paulus, F. M. (2020). Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nature Communications*, *11*(1), 1–13. https://doi.org/10.1038/s41467-020-14800-4

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210.

Thoits, P. A. (1995). Stress, coping, and social support processes: where are we? What next? *Journal of Health and Social Behavior*, pp. 53–79. https://doi.org/10.2307/2626957

van Leeuwen, J. M. C., Vink, M., Joëls, M., Kahn, R. S., Hermans, E. J., & Vinkers, C. H. (2019a). Increased responses of the reward circuitry to positive task feedback following acute stress in healthy controls but not in siblings of schizophrenia patients. *NeuroImage*, *184*, 547–554. https://doi.org/10.1016/j.neuroimage.2018.09.051

van Leeuwen, J. M. C., Vink, M., Joëls, M., Kahn, R. S., Hermans, E. J., & Vinkers, C. H. (2019b). Reward-related striatal responses following stress in healthy individuals and patients with bipolar disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(11), 966–974. https://doi.org/10.1016/j.bpsc.2019.06.014

vanDellen, M. R., Campbell, W. K., Hoyle, R. H., & Bradfield, E. K. (2011). Compensating, Resisting, and Breaking: A Meta-Analytic Examination of Reactions to Self-Esteem Threat. *Personality and Social Psychology Review*, *15*(1), 51–74. https://doi.org/10.1177/1088868310372950

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., & Gelman, A. (2019). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, R package version 2.1.0*.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural Circuitry of Reward Prediction Error. *Annual Review of Neuroscience*, *40*(1), 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Weitzman, E. D., Fukushima, D., Nogeire, C., Roffwarg, H., Gallagher, T. F., & Hellman, L. (1971). Twenty-four Hour Pattern of the Episodic Secretion of Cortisol in Normal Subjects. *The Journal of Clinical Endocrinology & Metabolism*, *33*(1), 14–22. https://doi.org/10.1210/jcem-33-1-14

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wolf, O. T., Minnebusch, D., & Daum, I. (2009). Stress impairs acquisition of delay eyeblink conditioning in men and women. *Neurobiology of Learning and Memory*, *91*(4), 431–436. https://doi.org/10.1016/j.nlm.2008.11.002

5.7 Supplementary Information

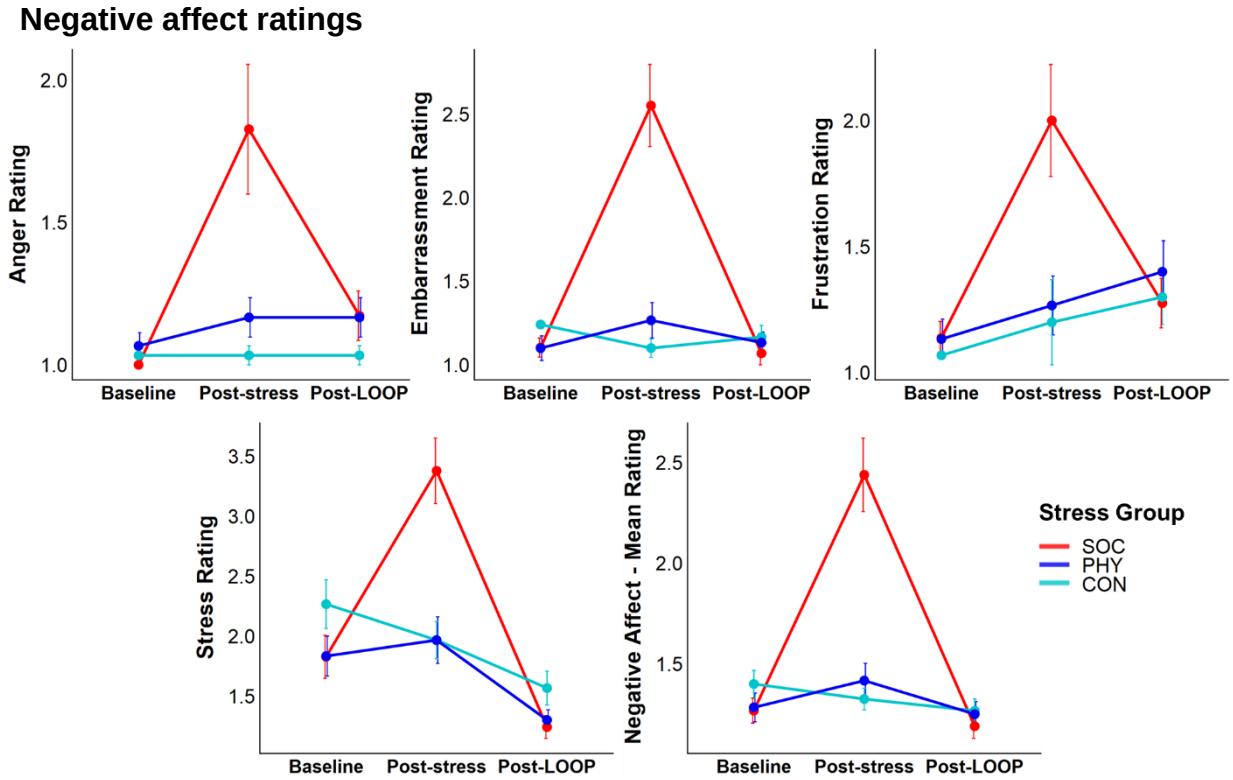
Supplementary Results

Cortisol response

Supplementary Table 5.1. Cortisol response - Scheirer-Ray-Hare Test

	Sum of Squares	*df*	*H*	*p*
Stress group	12642	2	18.939	< .001
Time of the day	6830	2	10.232	.006
Stress group x Time of the day	3640	4	5.902	.207
Residuals	35329	80		

Note. Group comparison of the stress-induced cortisol response (post-stress $T2_{CORT}$ - baseline T1); *df* = degrees of freedom; *H* = test statistic; factor Stress group: social-evaluative stress (*n* = 29) vs. physical stress (*n* = 30) vs. no stress (*n* = 30), factor time of the day: morning vs. noon vs. afternoon.



Supplementary Figure 5.1. Means and standard errors for the negative affect ratings separately for the three stress groups (SOC = social-evaluative stress [*n* = 29], PHY = physical stress [*n* = 30], CON =control [*n* = 30, embarrassment and frustration: *n* = 29 due to missing values]); LOOP = Learning of own performance task.

Forming self-related beliefs over time - Model free behavior analysis

Supplementary Table 5.2. Performance Expectation Ratings - Linear model

	B [95 % CI]	*SE*	*df*	*t*	*p*
Intercept	42.33 [40.48; 44.19]	0.95	5156	44.70	< .001
Ability condition	9.59 [7.36; 11.83]	1.13	86	8.52	<.001
Ability condition ✻ SOC vs. [PHY, CON]	-0.35 [-1.94; 1.24]	0.80	86	-0.43	.665
Ability condition ✻ PHY vs. CON	0.79 [-1.93; 3.52]	1.37	86	0.58	.564
Trial	-0.41 [-0.44; -0.37]	0.02	5156	-23.75	< .001
Trial ✻ SOC vs. [PHY, CON]	-0.01 [-0.04; 0.01]	0.01	5156	-0.96	.335
Trial ✻ PHY vs. CON	0.02 [-0.02; 0.06]	0.02	5156	1.13	.261
Trial ✻ Ability condition	0.80 [0.75; 0.84]	0.02	5156	32.73	< .001
Trial ✻ Ability condition ✻ SOC vs. [PHY, CON]	0.07 [0.04; 0.10]	0.02	5156	4.01	< .001
Trial ✻ Ability condition ✻ PHY vs. CON	-0.06 [-0.12; -0.01]	0.03	5156	-2.15	.031
SOC vs. [PHY, CON]	0.96 [-0.38; 2.30]	0.67	86	1.43	.157
PHY vs. CON	-0.48 [-2.77; 1.81]	1.15	86	-0.42	.677

Note. Linear mixed-effects model fit by maximum likelihood; dependent variable: performance expectation ratings; continuous variable: Trial, factor variables: Ability condition (high vs. low) and Stress group (SOC = social-evaluative stress [*n* = 29], PHY= physical stress [*n* = 30], CON = control [*n* = 30]) split in the contrasts SOC vs. [PHY,CON] and PHY vs. CON; *B* = unstandardized beta coefficient; *CI* = 95 % confidence interval; *SE* = standard error of B; *df* = degrees of freedom.

Model Selection

Supplementary Table 5.3. Model comparison

Model	*PSIS-LOO*	*LOO-SE*	*LOO-Diff (SE-Diff)*	*% of $\hat{k}$ > 0.7*	*No. Est. Parameters*
Whole Sample					
Unity Model	-2028.5	257.0	267.1 (52.0)	0.09	3
Ability Model	-1884.4	247.4	123.0 (95.9)	0.53	4
Valence Model	-1761.4	280.4		0.17	4
Mean Model	-2531.9	219.2	770.5 (93.5)	0.00	2
Social-evaluative Stress					
Unity Model	-625.3	83.1	60.7 (21.4)	0.17	3
Ability Model	-605.1	91.8	40.5 (16.4)	0.80	4
Valence Model	-564.6	91.7		0.29	4
Mean Model	-877.4	94.0	312.7 (40.3)	0.00	2
Physical Stress					
Unity Model	-840.1	225.7	107.6 (43.3)	0.00	3
Ability Model	-782.9	208.8	50.5 (62.1)	0.39	4
Valence Model	-732.5	247.5		0.11	4
Mean Model	-905.5	181.1	173.1 (75.1)	0.00	2
Control					
Unity Model	-563.1	92.2	98.7 (19.9)	0.11	3
Ability Model	-496.4	96.8	32.1 (17.3)	0.40	4
Valence Model	-464.3	98.3		0.11	4
Mean Model	-749.0	84.6	284.7 (35.3)	0.00	2

Note. LOO = sum PSIS-LOO, approximate leave-one-out cross-validation (LOO) using Pareto-smoothed importance sampling (PSIS); LOO-SE = Standard error of PSIS-LOO; LOO-Diff (SE-Diff) = Difference in expected predictive accuracy (PSIS-LOO) for all models from the model with the highest PSIS-LOO (Valence Model) and standard errors of differences; percentage of $\hat{k}$ - estimated shape parameters of the generalized Pareto distribution - exceeding 0.7 (all according to Vehtari et al., 2017); No. Est. Parameters = number of estimated parameters in the model; social-evaluative stress (n = 29), physical stress (n = 30), control (n = 29).

Posterior predictive checks: Behavioral analyses on the predicted data

Supplementary Table 5.4. Predicted Performance Expectations - Linear model

	B [95 % CI]	*SE*	*df*	*t*	*p*
Intercept	42.56 [40.80; 44.33]	0.90	5098	47.13	< .001
Ability condition	9.05 [7.01; 11.08]	1.03	85	8.82	<.001
Ability condition ✻ SOC vs. [PHY, CON]	-0.03 [-2.92; 2.86]	1.45	85	-0.02	.983
Ability condition ✻ PHY vs. CON	1.30 [-1.56; 4.17]	1.44	85	0.90	.369
Trial	-0.42 [-0.44; -0.41]	0.01	5098	-49.37	< .001
Trial ✻ SOC vs. [PHY, CON]	0.00 [-0.02; 0.03]	0.01	5098	0.40	.690
Trial ✻ PHY vs. CON	0.03 [0.01; 0.06]	0.01	5098	2.83	.005
Trial ✻ Ability condition	0.84 [0.82; 0.87]	0.01	5098	69.66	< .001
Trial ✻ Ability condition ✻ SOC vs. [PHY, CON]	0.07 [0.03; 0.10]	0.02	5098	3.93	< .001
Trial ✻ Ability condition ✻ PHY vs. CON	-0.10 [-0.13; -0.07]	0.02	5098	-5.85	< .001

Note. Linear mixed-effects model fit by maximum likelihood; dependent variable: performance expectations predicted by winning model; continuous variable: Trial, factor variables: Ability condition (high vs. low) and Stress group (SOC = social-evaluative stress [*n* = 29], PHY= physical stress [*n* = 30], CON = control [*n* = 29]) split in the contrasts SOC vs. [PHY,CON] and PHY vs. CON; *B* = unstandardized beta coefficient; *CI* = 95 % confidence interval; *SE* = standard error of B; *df* = degrees of freedom.

Learning parameters

Group comparison of learning rates

Supplementary Table 5.5. Learning rates - Linear model

	B [95 % CI]	*SE*	*b*	*df*	*t*	*p*
Intercept	0.091 [0.079; 0.105]	0.007		85	13.412	< 0.001
PE-Valence	-0.013 [-0.020; -0.006]	0.004	-0.178	85	-3.596	< .001
SOC vs. [PHY, CON]	0.015 [-0.004; 0.034]	0.010	0.138	85	1.500	.137
PHY vs. CON	-0.007 [-0.023; 0.010]	0.008	-0.074	85	-0.798	.427
PE-Valence ✻ SOC vs. [PHY, CON]	0.012 [0.002; 0.022]	0.005	0.114	85	2.303	.024
PE-Valence ✻ PHY vs. CON	-0.003 [-0.012; 0.006]	0.004	-0.036	85	-0.724	.471

Note. Linear mixed-effects model fit by maximum likelihood; dependent variable: learning rates derived from the valence model; learning rates for positive and negative prediction errors (PE, within subject factor PE-Valence); Stress group (SOC = social-evaluative stress [*n* = 29], PHY= physical stress [*n* = 30], CON = control [*n* = 29]) split in the contrasts SOC vs. [PHY,CON] and PHY vs. CON; *B* = unstandardized beta coefficient; *CI* = 95 % confidence interval; *SE* = standard error of *B*; *b* = standardized beta coefficient; *df* = degrees of freedom.

Associations of valence bias score with stress response.

As was to be expected, both measured components of the stress response, i.e. change in negative affect (ΔAFF, post-stress $T2_{AFF}$ - baseline $T1_{AFF}$) and the cortisol response (ΔCORT, post-stress $T2_{CORT}$ - baseline $T1_{CORT}$) share common variance ($\rho_{\Delta AFF,\Delta CORT}$ = .31, p = .003). In order to test the effect of one component on the valence bias score (BIAS, $(\alpha_{PE+} - \alpha_{PE-})/(\alpha_{PE+} + \alpha_{PE-})$) independently of the other, a combined rank regression BIAS ~ ΔAFF + ΔCORT + TIME was calculated additionally. Neither the change in negative affect nor the change in cortisol could predict the valence bias score independently of the other stress component ($b_{\Delta AFF}$ = 0.054, t_{85} = 0.960, p = .340, $b_{\Delta CORT}$ = 0.016, t_{85} = 1.683, p = .096).

Within the subsamples of the three stress groups neither change in negative affect nor cortisol change could predict the valence bias score (all beta weights of the rank regression BIAS ~ ΔAFF as well as BIAS ~ ΔCORT + TIME are not significant, Table S6).

Supplementary Table 5.6. Rank regression for the valence bias score predicted by the change in negative affect and cortisol change

		Social Stress		**Physical Stress**		**Control**	
		ΔAFF	**ΔCORT | TIME**	**ΔAFF**	**ΔCORT | TIME**	**ΔAFF**	**ΔCORT | TIME**
Valence Bias Score	*b*	.021	.003	.393	.030	-.199	.045
	p	.749	.789	.089	.200	.275	.067
	n	29	29	30	30	29	29

Note. Rank regression BIAS ~ ΔAFF and BIAS ~ ΔCORT + TIME. Valence bias score = $(\alpha_{PE+} - \alpha_{PE-})/(\alpha_{PE+} + \alpha_{PE-})$, ΔAFF = change in negative affect (post-stress $T2_{AFF}$ - baseline $T1_{AFF}$), ΔCORT = Cortisol change (post-stress $T2_{CORT}$ - baseline $T1_{CORT}$), TIME = time of the day (morning vs. noon vs. afternoon); *b* = beta-weight of rank regression.

Associations of valence bias score with affective recovery

Supplementary Table 5.7. Rank regression of the affective recovery predicted by the valence bias score controlled for initial change in affect and modulated by the social stress group

	b	*SE*	*t*	*p*
Intercept	0.140	0.055	2.561	.012
BIAS	0.238	0.124	1.922	.058
SOC vs. [PHY, CON]	0.124	0.081	1.534	.129
PHY vs. CON	-0.010	0.059	-1.171	.865
ΔAFF	0.816	0.069	11.831	< .001
BIAS ✻ SOC vs. [PHY, CON]	0.322	0.189	1.702	.093
BIAS ✻ PHY vs. CON	0.005	0.139	0.033	.973

Note. Rank regression for the whole sample (n = 88, Affective Recovery ~ BIAS + ΔAFF + SOC vs. [PHY, CON] + PHY vs. CON + BIAS✻ SOC vs. [PHY, CON] + BIAS✻ PHY vs. CON). Affective recovery = post-stress $T2_{AFF}$ - post-learning $T3_{AFF}$; valence bias score/ BIAS = $(\alpha_{PE+} - \alpha_{PE-})/(\alpha_{PE+} + \alpha_{PE-})$; ΔAFF = change in negative affect (post-stress $T2_{AFF}$ - baseline $T1_{AFF}$); *b* = beta-weight of rank regression.

Associations of valence bias score with cortisol recovery.

Supplementary Table 5.8. Rank regression of the cortisol recovery predicted by the valence bias score controlled for initial change in cortisol

	Cortisol recovery | ΔCORT											
	Social Stress			**Physical Stress**			**Control**			**Whole sample**		
Valence Bias Score	*b*	*p*	*n*	*b*	*p*	*n*	*b*	*p*	*n*	*b*	*p*	*n*
	-0.58	.706	29	0.61	.663	30	-1.31	.252	29	0.19	.804	88

Note. Rank regression Cortisol Recovery ~ BIAS + ΔCORT. Cortisol recovery = post-stress $T2_{CORT}$ - post-learning $T3_{CORT}$; Valence bias score = $(\alpha_{PE+} - \alpha_{PE-})/(\alpha_{PE+} + \alpha_{PE-})$; ΔCORT = stress-induced cortisol change (post-stress $T2_{CORT}$ - baseline $T1_{CORT}$); *b* = beta-weight of rank regression.

Supplementary Tables

Supplementary Table 5.9a. Sample characteristics

	Social Stress			Physical Stress			Control			Test	
	M	*Md*	*SD*	*M*	*Md*	*SD*	*M*	*Md*	*SD*	*H(2)*	*p*
Age	22.9	23	2.76	22.5	23	1.94	22.3	22	3.00	1.47	.480
Self-esteem	6.44	6.75	1.02	6.30	6.42	0.94	6.02	6.25	0.93	5.03	.080
SIAS	1.91	1.90	0.51	1.96	1.92	0.31	2.02	2.00	0.60	1.23	.540
Cortisol baseline	8.04	7.07	5.22	6.17	4.88	4.17	7.30	5.09	5.89	1.74	.419
Affective state baseline	1.27	1.25	0.33	1.28	1.25	0.39	1.40	1.25	0.39	3.21	.201

Note. Sample characteristics for the three stress groups. *M* = mean; *Md* = median; *SD* = standard deviation; self-esteem assessed via averaged scores of the Self-Description Questionnaire (SDQ-III); SIAS = averaged score on the Social Interaction Anxiety Scale; *H* = Kruskal-Wallis Chi-squared.

Supplementary Table 5.10b. Sample characteristics

		Social Stress	Physical Stress	Control	Test	
					H	*p*
Gender	**female**	21	20	20		
	male	8	10	10	0.3 (*df*=2)	.861
Time of day	**morning**	10	10	10		
	noon	11	10	8		
	afternoon	8	10	12	1.27 (*df*=4)	.867

Note. Frequency distribution for gender and time of day of the measurement for the three stress groups; *H* = Pearson's Chi-squared test statistic

Supplementary References

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

6 General Discussion

6.1 Summary of findings

The studies investigated how biases in self-belief formation relate to emotional experiences, neural processes, and psychiatric symptoms and how they differ by learning context. Across all studies, people tend to update their beliefs about their abilities more strongly in response to negative feedback than to positive feedback. This results in an overall negativity bias in forming self-beliefs. Those with lower self-esteem demonstrate a more pronounced negativity bias, linking their belief updates to pre-existing negative self-concepts. While there are no context effects of an audience during learning in general, higher subclinical social anxiety amplifies more negatively biased updating in public, suggesting an interaction between social context and anxiety. As these findings already suggest a link between the affective experience and biased belief updating, Study 2 could demonstrate that biased self-belief updating is associated with the emotions of embarrassment and pride, where more negative updates align with more embarrassment and less pride. Furthermore, biased self-belief updating is linked to pupil dilation response and neural response to prediction errors in the anterior insula, amygdala, ventral tegmental area/ substantia nigra, and mPFC. In a clinical sample with depression, the insula shows heightened sensitivity to negative self-related prediction errors, suggesting that depression involves a more intense response to negative feedback. The negativity bias in self-belief updating in individuals with depression does not differ from a healthy control group overall. However, a relationship between biased belief updating and symptom burden becomes apparent in the dimensional approach. Here, a higher symptom burden is linked to more negative updating of self-beliefs and more positive belief updating about others. Within the depressed group, greater symptom severity is related to less positive belief updates. Study 4 showed a reduced negativity bias after social-evaluative stress, suggesting that stress temporarily diminishes the focus on negative self-related information. This reduced negativity bias was linked to a decrease in stress-induced negative affect, potentially allowing for the regeneration of state self-worth.
Overall, the studies provide insights into how we arrive at our self-beliefs and reveal that cognitive biases in self-belief updating are intertwined with affective experiences. This is linked to individual factors like levels of mental health, which can further interact with context factors like publicity.

6.2 Negativity bias in self-belief updating

The consistent finding across all studies is a bias in forming self-beliefs that favors greater integration of negative feedback. This bias was found in different versions of the LOOP task: the one with self-related learning only (Study 1 Audience LOOP and Study 4 CPT and control group) as well as the version with additional other-related learning (Agent LOOP, Study 1 - 3). This demonstrates the self-specificity of the bias. A negative bias was also found in both public and private contexts during task execution (Study 1). Although a negativity bias is consistent with some studies using a variety of different paradigms

(Brotzeller & Gollwitzer, 2024; Ertac, 2011; Zamfir & Dayan, 2022), it contrasts with the dominant literature on a positive bias in updating self-beliefs (Sharot, 2011). This raises the question of how the LOOP task differs from other studies that consistently report a positivity bias. A key difference is that it is an active performance context with trial-by-trial task execution. This creates the impression among the participants that they can influence the feedback through their performance (even though it is controlled in the end). A task engagement with active performance can elicit specific self-related motivation. First, a motivation to improve one's ability from one trial to the next can be assumed (Jordan & Audia, 2012; Taylor et al., 1995). Situations that enable improvement tend to ignite a desire for self-improvement, particularly in response to previous failures (Taylor et al., 1995). For this, negative feedback is relevant, as it indicates that adaptation in behavior is needed (Sedikides & Hepper, 2009). The emotion of pride could support the presence of self-improvement motivation during task performance, as pride can reflect the perception of successful improvement (Tracy & Robins, 2007b). However, it was associated with less negatively biased belief updating. Also, when manipulating the opportunity for improvement, there is no effect on the negativity bias in self-belief updating. This suggests that negative feedback, even though relevant, is not incorporated into self-beliefs when motivated to improve (Brotzeller & Gollwitzer, 2024).

Second, active performance is also accompanied by the risk of failure. This can elicit the motivation to avoid failure, negative social evaluation, and embarrassment (Atkinson, 1957; Goffman, 2023). A fear of social evaluation can shift the focus toward negative feedback in the sense of threat monitoring (Shechner & Bar-Haim, 2016). This could lead to a stronger belief updating in the negative direction. The presence of embarrassment, which was also associated with biased belief updating, indicates that the participants perceived publicity and imagined social evaluation, further supporting an avoidance motivation. Additionally, increased response in pupil dilation was linked to more negative prediction errors, which could indicate increased arousal and attention to more negative prediction errors. Interestingly, this link was modulated by increased feelings of embarrassment, decreased feelings of pride, as well as a more negatively biased self-belief updating. As the negativity bias was also present in the private self-only version of the LOOP task, it may still have triggered a fear of failure through imagined publicity and evaluation by the experimental set-up in general, including the recording of response behavior. As both positive and negative affect were experienced during task execution, it can be assumed that both approach and avoidance motivations were present. An individually stronger focus on improvement or avoidance motivation could explain the inter-individual differences in the bias. As emotion can shift attention toward corresponding salient stimuli and intensify learning through arousal (Christianson, 2014), a stronger presence of either positive or negative affect can individually bias self-belief formation. As motivations and emotions relate to one's performance only, self-belief formation is biased, while the formation of beliefs about the other person is not.

Task characteristics and the negativity bias

A closer look at specific task characteristics can provide further insights into the background of the learning bias. First, the **valence** of self-belief that was addressed, a belief about one's estimation abilities, is usually a belief with a rather neutral, sometimes even negative expression. The starting values of the performance expectation ratings were usually around 50 % or slightly below. In contrast, paradigms that find a positive bias typically address self-beliefs that can be assumed to be rather positive (Korn et al., 2012; Möbius et al., 2011). In the present studies, more negative prior beliefs may have distorted the updating process negatively in a confirmatory way (Mokady & Reggev, 2022). As shown in Study 1, self-esteem is linked to the bias in belief formation, which supports the association with prior beliefs. Second, the **precision** of the self-beliefs addressed can impact the updating process (Bromberg-Martin & Sharot, 2020). For the LOOP task, a domain was selected where people's prior beliefs are usually vague (i.e., low precision). This allowed a relatively new belief to develop throughout the experiment and made the cover story not readily apparent. Low belief precision about one's ability usually arises from little prior experience. Thus, confidence in answering the estimation questions was probably low, accompanied by a general feeling of uncertainty during task performance (Pouget et al., 2016). It might have created a personal impression of being bad at the task. Studies show that, even without feedback, confidence when answering a task contributes to the formation of ability self-beliefs over time (Rouault et al., 2019). Thus, the negativity bias could reflect the personal perception of one's ability as feedback that corresponds to the subjective feeling appears more plausible and is preferentially integrated (Swann, 1983). Furthermore, when selecting the estimation questions and the corresponding answer scale, it was ensured that the correct answer was not identifiable so that pre-programmed feedback could be presented without jeopardizing the cover story. This also made the task more difficult and could have further contributed to low confidence in answering the estimation questions and, thus, the perception of generally low estimation abilities. Third, within the belief hierarchy from global to specific, a relatively **specific** self-belief was addressed (Marsh & Shavelson, 1985). This generally made the integration of negative feedback more likely, as the self is less threatened at this level of the hierarchy (Engerer et al., 2019; Hattie & Timperley, 2007). Whereby it should be noted here that tasks that address ability self-beliefs on a more global level also find a negativity bias (Brotzeller & Gollwitzer, 2024). Fourth, the **feedback** was relatively **precise** about one's performance score and **social-comparative**. Social comparison may have triggered performance rather than mastery goals (Elliot & McGregor, 2001), that is, a stronger focus on one's performance scores relative to others rather than improving one's skills. The precise feedback left little room for a self-serving interpretation in case of negative feedback, making it more threatening. This would rather point to performance-avoidance goals for this task environment and support a threat-driven focus on negative feedback. Assuming that both improvement motivation/ mastery goals and avoidance motivation/ performance-avoidance goals were present, as described above, the type of feedback tends to favor a stronger weighting on avoidance motivation. The feedback, consisting of

percentile ranks, also did not provide useful information like instructions for an actual improvement.

In order to investigate the specific influence of individual task characteristics (cf., Table 3), further experiments with targeted manipulation are required. For example, targeting different types of self-beliefs with varying valence and precision of prior beliefs in a within-subject design or specifically manipulating task engagement (active vs passive), task difficulty, or manipulation feedback precision (e.g., "With a probability of 75%, you are…"; cf. Möbius et al., 2011) can provide further insights.

Table 3. a. Overview of possible influences of task characteristics on self-belief updates in the LOOP task

Within the task			
Type of self-beliefs addressed			
Valence	Positive		**Negative**
Precision	High		**Low**
Position within hierarchy	Global		**Specific**
Type of feedback			
Valence	Positive PE		Negative PE
Precision	**High**		Low
Social	**Social-comparative**		Non-comparative
Task engagement	**Active performance**		Passive feedback
Task difficulty	**High**	Low	

Note. Task characteristics of the Learning of own performance (LOOP) task that potentially contributed to biased self-belief formation. Specifications of certain task characteristics that might be linked to the negativity bias are highlighted in bold blue. Grey = manipulated as an independent variable, PE = prediction error.

Negativity bias in different contexts

In Studies 1 and 4, social context factors were manipulated to investigate variation in biased self-belief updating under different conditions. Generally, a social context was already introduced through social-comparative feedback within the LOOP task. Furthermore, a second participant supposedly observed the participant's performance in the Agent-LOOP version of the task, which made the context slightly social-evaluative. To investigate how a social public context influences the formation of beliefs, it was varied in different stages. In Study 1, publicity was established in a minimal fashion in experiment three by allegedly making the performance rating phase public for the other participant. In experiment two, publicity was more pronounced by the experimenter sitting behind the participant throughout the task performance. Both manipulations did not show a

difference in self-belief formation. This could be due to manipulation involving only one passive observer, which may not have been strong enough to induce an impression of social evaluation that affects self-belief formation. A rapid habituation to the observer could also have been the case (Cacioppo et al., 1990). In case the induction of an audience effect worked, there also might have been a ceiling effect as the negativity bias was already present in the private self-only version of the LOOP task. It could also be considered that social presence and social evaluative threat are just not the driving force of the negativity bias, potentially favoring the explanation of high task difficulty and low confidence discussed above. However, in individuals prone to the fear of negative evaluation, the social context appeared to play a significant role, as social anxiety was only associated with a more negatively biased belief formation in public settings. This demonstrates the relevance of context characteristics in interaction with personality traits or symptom proneness. To better understand the effect of audience inductions, additional measures of physical arousal and affective reaction in public and private conditions could provide further insight, as well as additional induction of stronger forms of social publicity. In Study 4, a social-evaluative threat was induced more extremely by a public speech in front of a three-person audience. In contrast to an audience during task performance, the threat induction was applied before the LOOP task. As a result, the bias in self-belief formation shifted towards a less negative updating. This was associated with a better recovery from negative affect after the experienced self-threat. This further supports that belief updating is not solely driven by maximizing external outcomes but also internal states (Sharot et al., 2023). When internal states are affected by a prior threat induction, belief updating could have a function in itself, which is a better recovery from threatened state self-esteem and negative affect by a shift towards a more self-serving feedback processing. This is in line with increased self-enhancement following perceived failure (Hughes & Beer, 2013).

Table 2. b. Overview of manipulations of the context

Contextual embedding of the task/ experimental manipulation		
Social	Non-social	

- Presence of others <u>during</u> the task:
  - Agent-LOOP with other participant observing the feedback only (private, Studies 1 - 3) or feedback and expectation rating (public, Study 1)
  - Audience-LOOP with experimenter observing task performance (Study 1)
	Private self-only LOOP task (Studies 1 and 4)	

- Presence of others (threat induction) <u>before</u> the task		
Social evaluative threat (public speech)	CPT	No stress

Note. Bold blue = observed shift in negativity bias by social presence. CPT = cold pressure test

The value of a negatively biased self-belief

Negative self-beliefs and negatively biased belief updating are often discussed in contexts of mental disorders. Here, a number of studies have shown that negatively biased updating also exists in healthy samples. One could argue that the samples might have been slightly depressive or anxious or that negatively biased processing of performance feedback is a specific characteristic of young psychology students who were a dominating part of the participants. However, the negativity bias has been replicated in a variety of samples, even in older non-student samples (see control group of Study 3). One could further argue that maladaptive behavior also occurs in healthy people. Even if this is certainly true (and adaptive and maladaptive are not distinct categories anyway), it does not sufficiently address the negativity bias as an explanation. The results may support overcoming the dichotomy "positive equals healthy" and "negative equals pathological" often found in self-help narratives emphasizing positive thinking and sometimes also in research or cognitive behavioral therapy. Negative self-beliefs also have value. What makes a negative self-belief adaptive? At the global level, a majority of positive self-beliefs are certainly relevant for psychological well-being, while severe, rigid negative self-beliefs are likely associated with pathology (Beck, 2002). However, at lower levels of the belief hierarchy, a differentiated image of the self is useful (Linville, 1987). Exclusively positive self-beliefs about one's abilities are unrealistic and not informative when making decisions like in which activity to invest time and effort (e.g., which profession to learn or leisure activity to pursue). Generally, negative self-beliefs can support risk management, realistic self-appraisals, and self-improvement by promoting the practice of new skills, seeking additional support if needed, and preventing overconfidence (Colvin et al., 1995; Dunning et al., 2004; Leary, 2007). The motivation for self-accuracy, in particular, includes approaching negative self-beliefs to hold realistic beliefs about one's abilities (Sedikides & Strube, 1997). However, the presented results demonstrate a biased (i.e., inaccurate) self-belief formation in the negative direction. To understand the value of biased negative self-beliefs, it is important to consider the affective experience as well. The affectively-colored self-beliefs can predict where we "feel" good or bad at to guide us not only by external measures of ability but also by internal states. A negatively biased self-belief can be useful for keeping a distance from certain tasks to avoid negative emotions like frustration or embarrassment. However, the transition to maladaptive behavior is fluid. For example, pursuing a task can be helpful in the long term but frustrating in the short term. Frequently breaking off tasks or avoiding them due to a negatively biased self-belief can hinder the formation of global self-efficacy beliefs in the long run.

6.3 Neural underpinnings of the negativity bias in self-belief formation

Before having a closer look at the maladaptive aspects of biased self-formation, I will first focus on its neural underpinnings. Findings of Study 2 first identified brain regions within the mPFC and insula for the processing of prediction error surprise for both self-and other-related feedback and second, regions within the NAcc/VS, mPFC that specifically track prediction error valence for self-related but not other-related feedback. The processing of prediction errors was linked to individual levels of the biases in self-belief updating as well

as the current affective state. More precisely, responses to more negative prediction errors in the anterior insula, amygdala, mPFC, and VTA/ SN increased with more negative self-belief updating as well as more embarrassment and less pride. Similarly, a negative bias in pupil dilation, that is, a stronger pupil response to negative relative to positive prediction errors, modulated the neural prediction error response in the same way. Together, the stronger the neural response to more negative prediction errors, the greater the belief updating towards negative information as well as the negativity of the affective experience during task performance.

The self-specific results on the neural system level show the special emotional investment for self-related feedback, and it suggests that the incoming information, with its affective reaction and the corresponding arousal, are integrated to make a belief update. Especially the anterior insula has been discussed for integrating motivation, emotion, and cognition with context factors to guide behavior (Chang et al., 2013; Koban & Pourtois, 2014; Menon & Uddin, 2010; Wager & Barrett, 2017). Accordingly, activity in the anterior insula was associated with ratings of self-conscious emotions, physiological arousal (pupil dilation), as well as biased belief updates. The anterior insula has also been associated with the detection of salience, error awareness, and allocation of attention (Koban & Pourtois, 2014; Menon & Uddin, 2010; Touroutoglou et al., 2012; Ullsperger et al., 2010), making it potentially relevant for attentional shifts towards negative feedback in the sense of threat monitoring. Insula activity was also related to prediction error surprise, which may link the insula to uncertainty during task execution, another potential driver of the negativity bias (see section 6.2). Overall, the findings suggest the insula as a potential physiological basis for motivated and affectively colored updates of self-beliefs, resulting in the negativity bias.

The amygdala was similarly related to affect, pupil dilation, and belief updating bias in the LOOP task. This is in line with its link to error processing (Koban & Pourtois, 2014) and emotional learning, specifically learning from self-related negative feedback in a social context (Davis et al., 2010; Pejic et al., 2013). Also, in the VTA/SN and mPFC, the activity in response to prediction errors was modulated by belief updating bias and affect. This aligns with their relevance for prediction error processing and tracking subjective value. Especially the mPFC has also been linked to self-related information processing (Denny et al., 2012), the generation of affective meaning (Roy et al., 2012), and updating self-beliefs (Kuzmanovic et al., 2018; Will et al., 2017; Wittmann et al., 2016).

Psychophysiological interaction analyses revealed that the insula showed stronger functional connectivity with the amygdala, mPFC, and VTA/SN in response to negative prediction errors for self-related information compared to other-related information. This stronger self-related connectivity was modulated by biased belief formation. The functional connectivity findings support the insula's role in integrating context information, such as feedback, with affective and value-related information to update a belief. Together, the neural findings support the role of individual factors in the notable variability in biases of self-belief formation as prediction error signaling in brain regions critical for

affective and self-related information processing was modulated by individual levels of affective experience and arousal.

6.4 Neural and behavioral aspects of self-belief formation in depression

To better understand clinically relevant aspects of biased self-belief formation, a sample of individuals diagnosed with depression and variance in comorbid social anxiety was examined while performing the LOOP task. Complementing the neural results of Study 2, Study 3 found heightened insula reactivity to negative prediction errors in individuals with depression but no group difference in neural response to positive prediction errors. Insula reactivity to negative prediction errors was also associated with symptom burden. This is in line with a stronger insula reactivity to negative self-related feedback in individuals with depression (Jankowski et al., 2018; Kumar et al., 2017; Silk et al., 2014). It may reflect a stronger emotional reactivity to negative feedback as insula activity was linked to affective response in Studies 2 and 3. This would be in line with the diminished mean positive affect in individuals with depression during task performance. Heightened neural reactivity to negative social feedback has been linked to more negatively biased information processing (Jankowski et al., 2018). Receiving negative feedback also aligns with the frequent “emotional hotspots” in chronic depression, "making mistakes" (McCullough, 2003), which is typically associated with intense emotional reactions. This stronger neural and potentially emotional reactivity to negative feedback results in lower mood, which contributes to cognitive distortions like cognitive immunization (Kube & Korn, 2024). Together, the neural findings of Studies 2 and 3 illustrate a neurocomputational pathway in which affect and motivation bias self-belief updates with an enhanced reactivity to negative feedback.

Behaviorally, increased symptom burden was linked to more biased belief formation. More precisely, self-beliefs were more negative, while beliefs about others were more positive. In individuals with depression, in particular, updates in response to better-than-expected feedback decreased as their symptom burden increased. Both groups showed a self-specific negativity bias in belief updating without a difference at the group comparison level. As discussed above, the learning context, with rather low confidence in task performance and motivation to avoid failure, consistently leads to a negativity bias in belief updating, also in healthy individuals (see 6.2). Consequently, the shift from a healthy level of negative self-belief to a maladaptive one is quite subtle, making it plausible that symptom burden becomes apparent only in a dimensional analysis rather than through a straightforward group comparison.

More biased self-belief formation with increased symptom burden could be due to an influence of more negative global prior self-beliefs in the sample with depression. The link between more negatively biased belief formation and more negative global self-beliefs, indicated by a measure of self-esteem, has been shown before in Study 1. Also, the affective experience during task performance is linked to biased self-belief formation, as indicated by Study 2. Individuals with depression showed diminished happiness during task performance, potentially also due to an increased (emotional) response to negative self-related feedback, as mentioned above. Negative affect has been shown to hinder self-

belief updates following positive social feedback (Kube & Korn, 2024). The greater neglect of self-related positive prediction errors may result in more negative self-beliefs, especially in individuals with high levels of depression. It could result from the mechanism of cognitive immunization against positive information, that is, the cognitive devaluation of new evidence, which contributes to the maintenance of maladaptive self-beliefs in depression (Kube et al., 2019). It could also result from an attentional shift towards negative information, potentially due to heightened fear of negative evaluation and threat monitoring, which led to a neglect of positive information. Together, the interplay of more negative prior beliefs, more negative affect, and more negatively biased formation of novel beliefs confirming the negative global priors may be described in a vicious circle (Zamfir & Dayan, 2022) that perpetuates negative self-beliefs.

The findings of Study 1 also demonstrated the interplay between symptom expression and learning context, with increased negatively biased updates of self-beliefs in more socially anxious individuals only in a public context. Attentional shifts with monitoring failure and potential negative evaluation could drive the negativity bias. Once an evaluative threat is present, this might increase in those prone to fear of negative evaluation.

6.5 Clinical implications of biased self-belief formation

Understanding the formation of self-beliefs can provide insights into straining self-beliefs and corresponding emotions that patients report in therapeutic settings. As previously described, emotional or physical responses to stimuli, such as social feedback, can be understood as a reaction to the new evidence against the light of the prior belief (in Bayesian terms). This means the reaction is shaped by prior beliefs and the way new evidence is processed (Barrett, 2017). These prior beliefs are established over a history of learning, for example, repeated traumatic experiences. Consequently, an individual's current emotional reaction to a stimulus is influenced not only by the immediate context (e.g., a safe environment) but also by the learning history (e.g., "the world is dangerous"). Thus, current emotional responses are a blend of immediate reactions and cumulative responses to past evidence. As the precision on priors based on very aversive experiences is usually high and the likelihood of the new information comparatively low, a positive prediction error is hardly incorporated, which results in a posterior close to the prior (e.g., "This place is dangerous") and correspondingly, intensive emotional reactions. This layered response mirrors the concept of emotional schemata in psychotherapy, where past experiences and emotional patterns shape interpretations of new events (Greenberg, 2010). In the case of maladaptive emotional schemas formed through an aversive or even traumatic learning history, this can result in an extreme reaction to a less extreme event as the reaction is partly a reaction to a past event.

This framework, along with the findings, can offer insights into addressing maladaptive beliefs therapeutically. As several aspects are already present in existing psychotherapy practices, this discussion aims to bridge the gap between research results and therapeutic interventions. The findings may highlight why certain interventions are effective or suggest where a targeted focus could enhance therapeutic outcomes.

Looking at the reactivity to prediction errors first, the heightened sensitivity to negative feedback (see 6.3) could be validated through psychoeducation on emotions from a predictive processing perspective. Recognizing that events are processed through the lens of prior experiences can help contextualize one's current experience. This relates to identifying emotional schemas as done in emotion-focused therapy (Elliott et al., 2003) or schema therapy (Young et al., 2003). Generally, a common practice during early therapy sessions is collecting the predisposing, precipitating, and perpetuating factors in an individualized model of the disorder (Kuyken et al., 2011), which offers valuable insights into the learning history of self-beliefs in retrospect. Here, individual emotional hotspots such as “making mistakes” can be identified (McCullough, 2003), which can explain overly intense reactions to specific trigger stimuli that are linked to prior experiences. Next to contextualizing the emotional experience within the learning history, patients can be supported to get more detached from their prior beliefs in order to regulate emotional experience. Techniques such as defusion in Acceptance and Commitment Therapy (ACT; Hayes et al., 1999) or detached mindfulness in Metacognitive Therapy (MCT; Wells, 2011) can reduce patients' reactivity to entrenched negative self-concepts. As patients have difficulties integrating new positive evidence that challenges negative prior beliefs, especially with increasing symptom burden, it might not be successful just to cognitively challenge the belief by providing new evidence (e.g., with cognitive restructuring). Instead, it might be helpful to support patients in becoming more receptive to positive feedback while getting more detached from their prior beliefs (i.e., less confident in prior beliefs). Mindfulness skills in ACT or mindfulness-based cognitive therapy (Segal et al., 2018) may enhance patients' sensitivity to current experiences that challenge their priors. In case of a cognitive immunization against the new evidence (e.g., downplaying a positive event), therapists may mark the distortion with its symptom-maintaining effect and support reframing the new positive experiences. For threat monitoring of negative evaluations or other maladaptive prior-driven attention guidance, the Attention Training Technique within MCT (Knowles et al., 2016) can increase flexibility in shifting attention.

6.6 Conclusions

The studies of this thesis provide a new framework for studying the formation of self-beliefs and show a bias toward integrating negative feedback. It allows, first, a computational description of the belief formation process and, second, the investigation of links between process and individual as well as contextual factors. Computationally, the dynamics of self-belief formation can be best described by the agent the prediction error relates to (self- or other-related) and the valence of prediction errors (positive or negative). This consistently revealed a stronger weighting of prediction errors with negative valence only when forming self-beliefs. This self-related negativity bias in belief formation is associated with a variety of individual factors: first, the neural processing of prediction errors in the insula, amygdala, mPFC, and midbrain regions; second, the affective experience of more embarrassment and less pride, together with higher arousal as indicated by the pupil dilation, and third, global prior beliefs about the self as indicated by the self-esteem, biasing the belief formation process in a confirmatory way. These connections point

towards self-belief updating as a dynamic interplay between cognition and emotion. Contextually, social-evaluative threats can have an impact on self-belief formation, depending on the point in time. Self-belief formation under observation is more negatively biased in those prone to fear of negative evaluation. Self-belief formation following a social-evaluative stressor is less negatively biased with a regenerative effect from the stress-induced negative emotion. The role of social contexts, task characteristics, and individual differences in modulating the learning bias offers a more nuanced perspective on self-belief updating. It opens up practical approaches to specific task characteristics to be considered in further research to deepen the understanding of what influences a learning bias. The findings highlight the prevalence of negativity bias in self-belief updating also in healthy samples, which raises the question of its potential adaptive and maladaptive dimensions depending on certain contexts and psychological conditions. Clinically, the findings show a stronger neural reactivity to negative feedback in individuals with depression and point toward more biased self-belief formation with increased symptom burden. This supports the interplay between individual factors and learning context and further indicates a subtle fluid transition between adaptive and maladaptive formation of self-beliefs. The reduced integration of positive prediction errors with increased symptom burden in the sample with depression could point towards therapeutic interventions that promote a higher receptivity of positive contextual information and detachment of negative prior beliefs. Overall, this body of work provides a robust foundation for future investigations into the mechanisms of self-belief formation and updating and the translation to therapeutic practice.

Closing remarks

I hope that this work contributes to viewing self-beliefs as the result of a continuing learning process. We should bear in mind that this learning process lasts a lifetime, and our experiment is not detached from it. Looking at belief updating with a predictive processing perspective, this could mean that updating behavior seems more biased in the context of the experiment, and the participants' prediction is actually well founded in the entire learning history (and we just do not have full access to the priors). For example, when giving feedback on cancer risk, conservative updating regarding the individual risk can be valid with the background of previous experience (e.g., being cancer-free in the family for generations). We can speak of a bias in the sense of an unequal weighting of prediction errors within the experiment. However, we cannot imply that this is an “incorrect” belief updating. Therefore, strong claims that self-updating is always optimistically biased or negatively biased should also be handled with caution.

This work should contribute to a differentiated picture of self-updating. I hope it stimulates interest in the learning process behind self-beliefs and encourages people to consider it for a differentiated understanding of belief updating. We will neither be able to prevent people from entering an experiment with their individual learning history nor entirely control for it. However, I am convinced that the LOOP task shows a promising approach to studying a belief’s learning history in a controlled manner (Krach et al., 2024). It

addressed participants' prior learning history by deliberately choosing a domain with at least little previous experience for the task and including individual factors in the analyses. I hereby hope this work can contribute to a broadened understanding of self-beliefs as a function of learning history.

7 References

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. H. (Eds.). (1968). *Theories of cognitive consistency: a sourcebook*. https://psycnet.apa.org/fulltext/1968-35010-000.pdf

Adams, G. C., Balbuena, L., Meng, X., & Asmundson, G. J. G. (2016). When social anxiety and depression go together: A population study of comorbidity and associated consequences. *Journal of Affective Disorders*, *206*, 48–54.

Admon, R., & Pizzagalli, D. A. (2015). Dysfunctional reward processing in depression. *Current Opinion in Psychology*, *4*, 114–118.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*(3), 261–271.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, *80*(3), 260–267.

Apsler, R. (1975). Effects of embarrassment on behavior toward others. *Journal of Personality and Social Psychology*, *32*(1), 145–153.

Ashbaugh, A. R., Antony, M. M., McCabe, R. E., Schmidt, L. A., & Swinson, R. P. (2005). Self-Evaluative Biases in Social Anxiety. *Cognitive Therapy and Research*, *29*(4), 387–398.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64, Part 1*(6), 359–372.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bandura, A. (2001). Social cognitive theory: an agentic perspective. *Annual Review of Psychology*, *52*, 1–26.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *The American Psychologist*, *37*(2), 122–147.

Barnard, P., & Teasdale, J. (2014). *Affect, cognition and change: Re-modelling depressive thought*. Psychology Press. https://doi.org/10.4324/9781315804750

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(11), 1833.

Baumeister, R. F. (2019). The self. In E. J. Finkel & R. F. Baumeister (Eds.), *Advanced social psychology: The state of the science*. Oxford University Press.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529.

Beck, A. T. (1963). Thinking and depression: I. Idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, *9*(4), 324–333.

Beck, A. T. (1964). Thinking and depression: II. Theory and therapy. *Archives of General Psychiatry*, *10*(6), 561–571.

Beck, Aaron T. (1967). *Depression: Clinical, Experimental, and Theoretical Aspects*. Hoeber Medical Division, Harper & Row.

Beck, Aaron T. (1979). *Cognitive Therapy of Depression*. Guilford Press.

Beck, Aaron T. (2002). Cognitive models of depression. In R. L. Leahy & E. T. Dowd (Eds.), *Clinical Advances in Cognitive Psychotherapy: Theory and Application* (14 (1), pp. 29–61). Springer Publishing Company.

Beck, Aaron T., John Rush, A., Shaw, B. F., Emery, G., DeRubeis, R. J., & Hollon, S. D. (2024). *Cognitive therapy of depression*. Guilford Publications.

Blair, K., Geraci, M., Devido, J., McCaffrey, D., Chen, G., Vythilingam, M., Ng, P., Hollon, N., Jones, M., Blair, R. J. R., & Pine, D. S. (2008). Neural response to self- and other referential praise and criticism in generalized social phobia. *Archives of General Psychiatry*, *65*(10), 1176–1184.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607.

Brolsma, S. C. A., Vrijsen, J. N., Vassena, E., Rostami Kandroodi, M., Bergman, M. A., van Eijndhoven, P. F., Collard, R. M., den Ouden, H. E. M., Schene, A. H., & Cools, R. (2022). Challenging the negative learning bias hypothesis of depression: reversal learning in a naturalistic psychiatric sample. *Psychological Medicine*, *52*(2), 303–313.

Bromberg-Martin, E. S., & Sharot, T. (2020). The value of beliefs. *Neuron*, *106*(4), 561–565.

Brotzeller, F., & Gollwitzer, M. (2024). Exploring Asymmetries in Self-Concept Change After Discrepant Feedback. *Personality & Social Psychology Bulletin*, 1461672241232738.

Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education Principles Policy and Practice*, *22*(4), 444–457.

Button, K. S., Kounali, D., Stapinski, L., Rapee, R. M., Lewis, G., & Munafò, M. R. (2015). Fear of negative evaluation biases social evaluation inference: Evidence from a probabilistic learning task. *PloS One*, *10*(4), 1–15.

Cacioppo, J. T., Rourke, P. A., Marshall-Goodell, B. S., Tassinary, L. G., & Baron, R. S. (1990). Rudimentary physiological effects of mere observation. *Psychophysiology*, *27*(2), 177–186.

Canessa, N., Crespi, C., Motterlini, M., Baud-Bovy, G., Chierchia, G., Pantaleo, G., Tettamanti, M., & Cappa, S. F. (2013). The functional and structural neural basis of individual differences in loss aversion. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(36), 14307–14317.

Chang, L. J., Yarkoni, T., Khaw, M. W., & Sanfey, A. G. (2013). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cerebral Cortex (New York, N.Y.: 1991)*, *23*(3), 739–749.

Charpentier, C. J., De Neve, J.-E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of Affective Decision Making: How Do Feelings Predict Choice? *Psychological Science*, *27*(6), 763–775.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience and Biobehavioral Reviews*, *55*, 247–267.

Chen, J., Short, M., & Kemps, E. (2020). Interpretation bias in social anxiety: A systematic review and meta-analysis. *Journal of Affective Disorders*, *276*, 1119–1130.

Christianson, S. A. (2014). *The handbook of emotion and memory: Research and theory* (S.-A. Christianson, Ed.). Psychology Press. https://doi.org/10.4324/9781315807454

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(03), 181–204.

Clark, D. M., & Wells, A. (1995). A cognitive model of social phobia. In R.G. Heimberg, M. Liebowitz, D. Hope, F. Schneier (Ed.), *Social phobia: Diagnosis, assessment, and treatment* (pp. 69–93). Guilford.

Clark, J. E., Watson, S., & Friston, K. J. (2018). What is mood? A computational perspective. *Psychological Medicine*, *48*(14), 2277–2284.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology*, *68*(6), 1152–1162.

Craig, A. D. B. (2009). How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10*(1), 59–70.

Craig, A. D. Bud. (2009). How do you feel--now? The anterior insula and human awareness. *Nature Reviews. Neuroscience*, *10*(1), 59–70.

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7*(2), 189–195.

Crocker, J., & Park, L. E. (2004). The costly pursuit of self-esteem. *Psychological Bulletin*, *130*(3), 392–414.

Davis, F. C., Johnstone, T., Mazzulla, E. C., Oler, J. A., & Whalen, P. J. (2010). Regional response differences across the human amygdaloid complex during social conditioning. *Cerebral Cortex (New York, N.Y.: 1991)*, *20*(3), 612–621.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

Dearing, R. L., & Tangney, J. P. (2003). *Shame and Guilt*. Guilford Publications.

Deci, E. L., & Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, *11*(4), 227–268.

DePaulo, B. M., Epstein, J. A., & LeMay, C. S. (1990). Responses of the socially anxious to the prospect of interpersonal evaluation. *Journal of Personality*, *58*(4), 623–640.

Dickerson, S. S., Mycek, P. J., & Zaldivar, F. (2008). Negative social evaluation, but not mere social presence, elicits cortisol responses to a laboratory stressor task. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *27*(1), 116–121.

Diederen, K. M. J., Spencer, T., Vestergaard, M. D., Fletcher, P. C., & Schultz, W. (2016). Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron*, *90*(5), 1127–1138.

Dixon, M. L., & Gross, J. J. (2021). Dynamic network organization of the self: implications for affective experience. *Current Opinion in Behavioral Sciences*, *39*, 1–9.

Dobson, K. S., & Dozois, D. J. A. (2021). *Handbook of Cognitive-Behavioral Therapies, Fourth Edition*. Guilford Publications.

Dozois, D. J. A., & Beck, A. T. (2008). Chapter 6 - Cognitive Schemas, Beliefs and Assumptions. In K. S. Dobson & D. J. A. Dozois (Eds.), *Risk Factors in Depression* (pp. 119–143). Elsevier.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *5*(3), 69–106.

Eil, D., & Rao, J. M. (2010). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science (New York, N.Y.)*, *302*(5643), 290–292.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200.

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*, 6149.

Elder, J., Davis, T., & Hughes, B. L. (2022). Learning about the self: Motives for coherence and positivity constrain learning from self-relevant social feedback. *Psychological Science*, *33*(4), 629–647.

Elliot, A. J., & McGregor, H. A. (2001). A 2 X 2 achievement goal framework. *Journal of Personality and Social Psychology*, *80*(3), 501–519.

Elliot, Andrew J., Gable, S. L., & Mapes, R. R. (2006). Approach and avoidance motivation in the social domain. *Personality & Social Psychology Bulletin*, *32*(3), 378–391.

Elliot, Andrew J., & Thrash, T. M. (2004). The intergenerational transmission of fear of failure. *Personality & Social Psychology Bulletin*, *30*(8), 957–971.

Elliott, R., Watson, J. C., Goldman, R. N., & Greenberg, L. S. (2003). *Learning emotion-focused therapy: The process-experiential approach to change*. American Psychological Association.

Engerer, C., Berberat, P. O., Dinkel, A., Rudolph, B., Sattel, H., & Wuensch, A. (2019). Specific feedback makes medical students better communicators. *BMC Medical Education*, *19*(1), 51.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545.

Etkin, A., & Wager, T. D. (2007). Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *The American Journal of Psychiatry*, *164*(10), 1476–1488.

Everaert, J., Bronstein, M. V., Cannon, T. D., & Joormann, J. (2018). Looking through tinted glasses: depression and social anxiety are related to both interpretation biases and inflexible negative interpretations. *Clinical Psychological Science*, *6*(4), 517–528.

Everaert, J., Bronstein, M. V., Castro, A. A., Cannon, T. D., & Joormann, J. (2020). When negative interpretations persist, positive emotions don't! Inflexible negative interpretations encourage depression and social anxiety by dampening positive emotions. *Behaviour Research and Therapy*, *124*, 103510.

Faustino, B., Vasco, A. B., Silva, A. N., & Marques, T. (2020). Relationships between Emotional Schemas, Mindfulness, Self-Compassion and Unconditional Self-Acceptance on the Regulation of Psychological Needs. *Research in Psychotherapy (Milano)*, *23*(2), 442.

Feinberg, M., Willer, R., & Keltner, D. (2012). Flustered and faithful: embarrassment as a signal of prosociality. *Journal of Personality and Social Psychology*, *102*(1), 81–97.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *The American Psychologist*, *56*(3), 218–226.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1431), 459–473.

Frolichs, K. M. M., Rosenblau, G., & Korn, C. W. (2022). Incorporating social knowledge structures into computational models. *Nature Communications*, *13*(1), 6205.

Gable, S. L. (2006). Approach and avoidance social motives and goals. *Journal of Personality*, *74*(1), 175–222.

Garrett, N., González-Garzón, A., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating Beliefs Under Perceived Threat. *Journal of Neuroscience*. https://doi.org/10.2139/ssrn.3155415

Garrett, N., Sharot, T., Faulkner, P., Korn, C. W., Roiser, J. P., & Dolan, R. J. (2014). Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Frontiers in Human Neuroscience*, *8*, 639.

Globig, L. K., Blain, B., & Sharot, T. (2022). Perceptions of personal and public risk: Dissociable effects on behavior and well-being. *Journal of Risk and Uncertainty*, *64*(2), 213–234.

Goffman, E. (2023). The presentation of self in everyday life. In *Social Theory Re-Wired* (pp. 450–459). Routledge.

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., Reid, I., Hall, J., & Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain: A Journal of Neurology*, *134*(6), 1751–1764.

Greenberg, L. S. (2010). Emotion-Focused Therapy: A Clinical Synthesis. *FOCUS*, *8*(1), 32–42.

Gruenewald, T. L., Kemeny, M. E., Aziz, N., & Fahey, J. L. (2004). Acute threat to the social self: shame, social self-esteem, and cortisol activity. *Psychosomatic Medicine*, *66*(6), 915–924.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(22), 5623–5630.

Harlé, K. M., Chang, L. J., van 't Wout, M., & Sanfey, A. G. (2012). The neural mechanisms of affect infusion in social economic decision-making: a mediating role of the anterior insula. *NeuroImage*, *61*(1), 32–40.

Harter, S. (1999). *The construction of the self: A developmental perspective*. Guilford Press.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hayes, S., Strosahl, K., & Wilson, K. (1999). *Acceptance and commitment therapy* (Vol. 6). New York: Guilford press.

He, Z., Ao, X., Muhlert, N., Elliott, R., & Zhang, D. (2020). Neural substrates of expectancy violation associated with social feedback in individuals with subthreshold depression. *Psychological Medicine*, 1–9.

Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2014). Chapter 24 - A Cognitive-Behavioral Model of Social Anxiety Disorder. In S. G. Hofmann & P. M. DiBartolo (Eds.), *Social Anxiety (Third Edition)* (pp. 705–728). Academic Press.

Heitmann, C. Y., Peterburs, J., Mothes-Lasch, M., Hallfarth, M. C., Böhme, S., Miltner, W. H. R., & Straube, T. (2014). Neural correlates of anticipation and processing of performance feedback in social anxiety. *Human Brain Mapping*, *35*(12), 6023–6031.

Hepper, E. G., Gramzow, R. H., & Sedikides, C. (2010). Individual differences in self-enhancement and self-protection strategies: an integrative analysis. *Journal of Personality*, *78*(2), 781–814.

Hirsch, C. R., Clark, D. M., Mathews, A., & Williams, R. (2003). Self-images play a causal role in social phobia. *Behaviour Research and Therapy*, *41*(8), 909–921.

Hoefler, A., Athenstaedt, U., Corcoran, K., Ebner, F., & Ischebeck, A. (2015). Coping with self-threat and the evaluation of self-related traits: An fMRI study. *PloS One*, *10*(9), e0136027.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hopkins, A. K., Dolan, R., Button, K. S., & Moutoussis, M. (2021). A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious Individuals. *Computational Psychiatry (Cambridge, Mass.)*, *5*(1), 21–37.

Hsu, D. T., Sanford, B. J., Meyers, K. K., Love, T. M., Hazlett, K. E., Walker, S. J., Mickey, B. J., Koeppe, R. A., Langenecker, S. A., & Zubieta, J.-K. (2015). It still hurts: altered endogenous opioid activity in the brain during social rejection and acceptance in major depressive disorder. *Molecular Psychiatry*, *20*(2), 193–200.

Hsu, D. T., Sanford, B. J., Meyers, K. K., Love, T. M., Hazlett, K. E., Wang, H., Ni, L., Walker, S. J., Mickey, B. J., Korycinski, S. T., Koeppe, R. A., Crocker, J. K., Langenecker, S. A., & Zubieta, J.-K. (2013). Response of the μ-opioid system to social rejection and acceptance. *Molecular Psychiatry*, *18*(11), 1211–1217.

Hughes, B. L., & Beer, J. S. (2013). Protecting the Self: The Effect of Social-evaluative Threat on Neural Representations of Self. *Journal of Cognitive Neuroscience*, *25*(4), 613–622.

Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19*(2), 62–64.

Jankowski, K. F., Batres, J., Scott, H., Smyda, G., Pfeifer, J. H., & Quevedo, K. (2018). Feeling left out: depressed adolescents may atypically recruit emotional salience and regulation networks during social exclusion. *Social Cognitive and Affective Neuroscience*, *13*(8), 863–876.

Jenkins, J. M., & Oatley, K. (1996). Emotional episodes and emotionality through the life span. In *Handbook of Emotion, Adult Development, and Aging* (pp. 421–441). Elsevier.

Jordan, A. H., & Audia, P. G. (2012). Self-enhancement and learning from performance feedback. *Academy of Management Review. Academy of Management*, *37*(2), 211–231.

Jug, R., Jiang, X. S., & Bean, S. M. (2019). Giving and receiving effective feedback: A review article and how-to guide. *Archives of Pathology & Laboratory Medicine*, *143*(2), 244–250.

Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: consensus and controversy. *Neuron*, *63*(6), 733–745.

Karnick, A. T., Bauer, B. W., & Capron, D. W. (2024). Negative mood and optimism bias: An experimental investigation of sadness and belief updating. *Journal of Behavior Therapy and Experimental Psychiatry*, 101910.

Kelly, C., Toro, R., Di Martino, A., Cox, C. L., Bellec, P., Castellanos, F. X., & Milham, M. P. (2012). A convergent functional architecture of the insula emerges across imaging modalities. *NeuroImage*, *61*(4), 1129–1142.

Kessler, R. C., Stang, P., Wittchen, H. U., Stein, M., & Walters, E. E. (1999). Lifetime co-morbidities between social phobia and mood disorders in the US National Comorbidity Survey. *Psychological Medicine*, *29*(3), 555–567.

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test'-a tool for investigating psychobiological stress responses in a laboratory setting. In *Neuropsychobiology* (Vol. 28, Issues 1–2, pp. 76–81). https://doi.org/10.1159/000119004

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

Knowles, M. M., Foden, P., El-Deredy, W., & Wells, A. (2016). A systematic review of efficacy of the Attention Training Technique in clinical and nonclinical samples. *Journal of Clinical Psychology*, *72*(10), 999–1025.

Koban, L., Andrews-Hanna, J. R., Ives, L., Wager, T. D., & Arch, J. J. (2023). Brain mediators of biased social learning of self-perception in social anxiety disorder. *Translational Psychiatry*, *13*(1), 292.

Koban, L., & Pourtois, G. (2014). Brain systems underlying the affective and social monitoring of actions: an integrative review. *Neuroscience and Biobehavioral Reviews*, *46 Pt 1*(1), 71–84.

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., Wager, T. D., & Arch, J. J. (2017). Social Anxiety is Characterized by Biased Learning About Performance and the Self. *Emotion* . https://doi.org/10.1037/emo0000296

Koenig, S., Uengoer, M., & Lachnit, H. (2018). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology*, *55*(4), 1–12.

Kopala-Sibley, D. C., & Klein, D. N. (2017). Depressive Disorders: Presentation, Classification, Developmental Trajectories, and Course. In N. L. Cohen (Ed.), *Public health perspectives on depressive disorders.* (pp. 13–39). Johns Hopkins University Press.

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844.

Korn, C. W., Sharot, T., Walter, H., Heekeren, H. R., & Dolan, R. J. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, *44*(3), 579–592.

Krach, S., Müller-Pinzler, L., Czekalla, N., Schröder, A., Lübber, F., Rademacher, L., Stolz, D. S., Paulus, F. M., Wilhelm, I., & Mayer, A. V. (2024). Examining self-belief formation through artificial beliefs. In *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/2y5tv

Kube, T. (2023). Biased belief updating in depression. *Clinical Psychology Review*, *103*, 102298.

Kube, T., & Korn, C. (2024). Induced negative affect hinders self-referential belief updating in response to social feedback. *Emotion (Washington, D.C.)*. https://doi.org/10.1037/emo0001426

Kube, T., Rief, W., Gollwitzer, M., Gärtner, T., & Glombiewski, J. A. (2019). Why dysfunctional expectations in depression persist - Results from two experimental studies investigating cognitive immunization. *Psychological Medicine*, *49*(9), 1532–1544.

Kumar, P., Waiter, G., Ahearn, T., Milders, M., Reid, I., & Steele, J. D. (2008). Abnormal temporal difference reward-learning signals in major depression. *Brain: A Journal of Neurology*, *131*(Pt 8), 2084–2093.

Kumar, Poornima, Goer, F., Murray, L., Dillon, D. G., Beltzer, M. L., Cohen, A. L., Brooks, N. H., & Pizzagalli, D. A. (2018). Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *43*(7), 1581–1588.

Kumar, Poornima, Waiter, G. D., Dubois, M., Milders, M., Reid, I., & Steele, J. D. (2017). Increased neural response to social rejection in major depression. *Depression and Anxiety*, *34*(11), 1049–1056.

Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder. *Neuroscience and Biobehavioral Reviews*, *69*, 313–332.

Kurman, J. (2006). Self-enhancement, self-regulation and self-improvement following failures. *The British Journal of Social Psychology / the British Psychological Society*, *45*(Pt 2), 339–356.

Kurth, F., Zilles, K., Fox, P. T., Laird, A. R., & Eickhoff, S. B. (2010). A link between the systems: functional differentiation and integration within the human insula revealed by meta-analysis. *Brain Structure & Function*, *214*(5–6), 519–534.

Kuyken, W., Padesky, C. A., & Dudley, R. (2011). *Collaborative case conceptualization: Working effectively with clients in cognitive-behavioral therapy*. Guilford Publications.

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162.

Kuzmanovic, B., & Rigoux, L. (2017). Valence-Dependent Belief Updating: Computational Validation. *Frontiers in Psychology*, *8*(JUN), 1087.

Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). Influence of vmPFC on dmPFC Predicts Valence-Guided Belief Formation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(37), 7996–8010.

Laurenceau, J.-P., Barrett, L. F., & Pietromonaco, P. R. (1998). Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of Personality and Social Psychology*, *74*(5), 1238–1251.

Leary, M. R. (2007). Motivational and Emotional Aspects of the Self. *Annual Review of Psychology*, *58*(1), 317–344.

Leary, M. R., & Atherton, S. C. (1986). Self-Efficacy, Social Anxiety, and Inhibition in Interpersonal Encounters. *Journal of Social and Clinical Psychology*, *4*(3), 256–267.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, *107*(1), 34–47.

Leknes, S., & Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews. Neuroscience*, *9*(4), 314–320.

Leong, Y. C., Hughes, B. L., Wang, Y., & Zaki, J. (2019). Neurocomputational mechanisms underlying motivated seeing. *Nature Human Behaviour*, *3*(9), 962–973.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. SAGE Publications.

Lewinsohn, P. M. (1974). A behavioral approach to depression. *Essential Papers on Depression*, 150–172.

Lewis, M. (2008). Self-conscious emotions: Embarrassment, pride, shame, and guilt. *Handbook of Emotions*. https://psycnet.apa.org/record/2008-07784-046

Linville, P. W. (1987). Self-complexity as a cognitive buffer against stress-related illness and depression. *Journal of Personality and Social Psychology*, *52*(4), 663–676.

Lockwood, P. L., Apps, M. A. J., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(35), 9763–9768.

Lockwood, P. L., & Klein-Flügge, M. C. (2020). Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Social Cognitive and Affective Neuroscience*, *16*(8), 761–771.

Lockwood, P. L., & Wittmann, M. K. (2018). Ventral anterior cingulate cortex and social decision-making. *Neuroscience and Biobehavioral Reviews*, *92*, 187–191.

Loewenstein, G. (2006). Social science. The pleasures and pains of information [Review of *Social science. The pleasures and pains of information*]. *Science*, *312*(5774), 704–706.

London, M. (2003). *Job feedback: Giving, seeking, and using feedback for performance improvement*. Psychology Press. https://doi.org/10.4324/9781410608871/job-feedback-manuel-london

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.

Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology. General*, *105*(1), 3–46.

Mandelli, L., Petrelli, C., & Serretti, A. (2015). The role of specific early trauma in adult depression: A meta-analysis of published literature. Childhood trauma and adult depression. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *30*(6), 665–680.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, *62*(1), 103–134.

Markus, H., & Wurf, E. (1987). The Dynamic Self-Concept: A Social Psychological Perspective. *Annual Review of Psychology*, *38*(1), 299–337.

Markus, Hazel, & Nurius, P. (1986). Possible selves. *The American Psychologist*, *41*(9), 954–969.

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, *20*(3), 107–123.

Masten, C. L., Eisenberger, N. I., Borofsky, L. A., McNealy, K., Pfeifer, J. H., & Dapretto, M. (2011). Subgenual anterior cingulate responses to peer rejection: a marker of adolescents' risk for depression. *Development and Psychopathology*, *23*(1), 283–292.

Masten, C. L., Eisenberger, N. I., Borofsky, L. A., Pfeifer, J. H., McNealy, K., Mazziotta, J. C., & Dapretto, M. (2009). Neural correlates of social exclusion during adolescence: understanding the distress of peer rejection. *Social Cognitive and Affective Neuroscience*, *4*(2), 143–157.

McCullough, J. P., Jr. (2003). Treatment for chronic depression using Cognitive Behavioral Analysis System of Psychotherapy (CBASP). *Journal of Clinical Psychology*, *59*(8), 833–846.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure & Function*, *214*(5–6), 655–667.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225.

Miller, R. S. (1996). *Embarrassment: Poise and peril in everyday life*. Guilford Press.

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing self-confidence: Theory and experimental evidence. *Working Paper Series // Federal Reserve Bank of Boston*, No. 11-14.

Mokady, A., & Reggev, N. (2022). The Role of Predictions, Their Confirmation, and Reward in Maintaining the Self-Concept. *Frontiers in Human Neuroscience*, *16*, 824085.

Morrison, A. S., & Heimberg, R. G. (2013). Social anxiety and social anxiety disorder. *Annual Review of Clinical Psychology*, *9*(1), 249–274.

Morvan, C., & O'Connor, A. J. (2017). *A Theory of Cognitive Dissonance* (1st Edition). Macat International. https://doi.org/10.4324/9781912282432

Müller-Pinzler, L., Gazzola, V., Keysers, C., Sommer, J., Jansen, A., Frässle, S., Einhäuser, W., Paulus, F. M., & Krach, S. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, *119*, 252–261.

Müller-Pinzler, Laura, Krach, S., Krämer, U. M., & Paulus, F. M. (2017). The social neuroscience of interpersonal emotions. *Current Topics in Behavioral Neurosciences*, *30*, 241–256.

Mumford, D. (1992). On the computational architecture of the neocortex: II The role of cortico-cortical loops. *Biological Cybernetics*, *66*(3), 241–251.

Murray, E. A. (2007). The amygdala, reward and emotion. *Trends in Cognitive Sciences*, *11*(11), 489–497.

Näätänen, R., Paavilainen, P., Tiitinen, H., Jiang, D., & Alho, K. (1993). Attention and mismatch negativity. *Psychophysiology*, *30*(5), 436–450.

Nave, K., Deane, G., Miller, M., & Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisciplinary Reviews. Cognitive Science*, *11*(6), e1542.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, *2*(2), 175–220.

O'Banion, K., & Arkowitz, H. (1977). Social anxiety and selective memory for affective information about the self. *Social Behavior and Personality: An International Journal*, *5*(2), 321–328.

Pavlov, P. I., & Anrep, G. V. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. *Oxford University Press*. https://doi.org/10.5214/ans.0972-7531.1017309

Pejic, T., Hermann, A., Vaitl, D., & Stark, R. (2013). Social anxiety modulates amygdala activation during social conditioning. *Social Cognitive and Affective Neuroscience*, *8*(3), 267–276.

Peterburs, J., Sandrock, C., Miltner, W. H. R., & Straube, T. (2016). Look who's judging—Feedback source modulates brain activation to performance feedback in social anxiety. *NeuroImage*, *133*, 430–437.

Phelps, E. A. (2006). Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53.

Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: the need to belong and enhanced sensitivity to social cues. *Personality & Social Psychology Bulletin*, *30*(9), 1095–1107.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374.

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Frontiers in Neuroscience*, *5*(September), 115.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Richard, E., & Petty, J. A. (2014). *Attitude strength: Antecedents and consequences* (R. E. Petty & J. A. Krosnick, Eds.; 1st Edition). Psychology Press. https://doi.org/10.4324/9781315807041

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., & Thorwart, A. (2015). Expectancies as core features of mental disorders. *Current Opinion in Psychiatry*, *28*(5), 378–385.

Robinson, O. J., Cools, R., Carlisi, C. O., Sahakian, B. J., & Drevets, W. C. (2012). Ventral striatum response during reward and punishment reversal learning in unmedicated major depressive disorder. *The American Journal of Psychiatry*, *169*(2), 152–159.

Rothkirch, M., Tonn, J., Köhler, S., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain: A Journal of Neurology*, *140*(4), 1147–1157.

Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, *10*(1), 1–11.

Rouhani, N., & Niv, Y. (2019). Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, *236*(8), 2425–2435.

Roy, M., Shohamy, D., & Wager, T. D. (2012). Ventromedial prefrontal-subcortical systems and the generation of affective meaning. *Trends in Cognitive Sciences*, *16*(3), 147–156.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews. Neuroscience*, *15*(8), 549–562.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning internal representations by error-propagation*. *1*(6088), 318–362.

Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., Lam, J., Skandali, N., Siegel, J. Z., Ousdal, O. T., Prabhu, G., Dayan, P., Fonagy, P., & Dolan, R. J. (2017). Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry* , *74*(8), 790–797.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78.

Safra, L., Chevallier, C., & Palminteri, S. (2019). Depressive symptoms are associated with blunted reward learning in social contexts. *PLoS Computational Biology*, *15*(7), e1007224.

Samoilov, A., & Goldfried, M. R. (2000). Role of emotion in cognitive-behavior therapy. *Clinical Psychology: A Publication of the Division of Clinical Psychology of the American Psychological Association*, *7*(4), 373–385.

Schlenker, B. R., & Leary, M. R. (1982). Social anxiety and self-presentation: a conceptualization and model. *Psychological Bulletin*, *92*(3), 641–669.

Schlund, M. W., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Cataldo, M. F., Forbes, E. E., Dahl, R. E., & Ryan, N. D. (2010). Nothing to fear? Neural systems supporting avoidance behavior in healthy youths. *NeuroImage*, *52*(2), 710–719.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Sedikides, C., & Gregg, A. P. (2008). *Self-Enhancement Food for Thought*. *3*(2), 102–116.

Sedikides, C., & Hepper, E. G. D. (2009). Self-improvement. *Social and Personality Psychology Compass*, *3*(6), 899–917.

Sedikides, C., & Strube, M. J. (1997). Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better. In *Advances in Experimental Social Psychology* (Vol. 29, pp. 209–269). Elsevier.

Segal, Z., Williams, M., & Teasdale, J. (2018). *Mindfulness-Based Cognitive Therapy for Depression, Second Edition*. Guilford Publications.

Sharot, T. (2011). The optimism bias. In *Current Biology* (Vol. 21, Issue 23, pp. R941–R945). Cell Press. https://doi.org/10.1016/j.cub.2011.10.030

Sharot, T., & Garrett, N. (2016). Forming Beliefs: Why Valence Matters. *Trends in Cognitive Sciences*, *20*(1), 25–33.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479.

Sharot, T., Rollwage, M., Sunstein, C. R., & Fleming, S. M. (2023). Why and When Beliefs Change. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *18*(1), 142–151.

Shechner, T., & Bar-Haim, Y. (2016). Threat Monitoring and Attention-Bias Modification in Anxiety and Stress-Related Disorders. *Current Directions in Psychological Science*, *25*(6), 431–437.

Silk, J. S., Siegle, G. J., Lee, K. H., Nelson, E. E., Stroud, L. R., & Dahl, R. E. (2014). Increased neural response to peer rejection associated with adolescent depression and pubertal development. *Social Cognitive and Affective Neuroscience*, *9*(11), 1798–1807.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13*(8), 334–340.

Sliz, D., & Hayley, S. (2012). Major depressive disorder and alterations in insular cortical activity: a review of current functional magnetic imaging research. *Frontiers in Human Neuroscience*, *6*, 323.

Smart Richman, L., & Leary, M. R. (2009). Reactions to discrimination, stigmatization, ostracism, and other forms of interpersonal rejection: a multimotive model. *Psychological Review*, *116*(2), 365–383.

Smith, R. E., & Sarason, I. G. (1975). Social anxiety and the evaluation of negative interpersonal feedback. *Journal of Consulting and Clinical Psychology*, *43*(3), 429.

Somerville, L. H., Kelley, W. M., & Heatherton, T. F. (2010). Self-esteem modulates medial prefrontal cortical responses to evaluative social feedback. *Cerebral Cortex (New York, N.Y.: 1991)*, *20*(12), 3005–3013.

Späti, J., Chumbley, J., Brakowski, J., Dörig, N., Grosse Holtforth, M., Seifritz, E., & Spinelli, S. (2014). Functional lateralization of the anterior insula during feedback processing. *Human Brain Mapping*, *35*(9), 4428–4439.

Sperduti, M., Delaveau, P., Fossati, P., & Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: A brief review and meta-analysis. *Brain Structure and Function*, *216*(2), 151–157.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In *Advances in Experimental Social Psychology* (Vol. 21, pp. 261–302). Elsevier.

Steinmetz, J., Xu, Q., Fishbach, A., & Zhang, Y. (2016). Being observed magnifies action. *Journal of Personality and Social Psychology*, *111*(6), 852–865.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S. M., Haker, H., Seth, A. K., & Petzschner, F. H. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, *10*, 550.

Stolz, D. S., Müller-Pinzler, L., Krach, S., & Paulus, F. M. (2020). Internal control beliefs shape positive affect and associated neural dynamics during outcome valuation. *Nature Communications 2020 11:1*, *11*(1), 1–13.

Strachman, A., & Gable, S. L. (2006). What you want (and do not want) affects what you see (and do not see): avoidance social goals and social events. *Personality & Social Psychology Bulletin*, *32*(11), 1446–1458.

Straube, T., Kolassa, I.-T., Glauer, M., Mentzel, H.-J., & Miltner, W. H. R. (2004). Effect of task conditions on brain responses to threatening faces in social phobics: an event-related functional magnetic resonance imaging study. *Biological Psychiatry*, *56*(12), 921–930.

Straube, T., Mentzel, H.-J., & Miltner, W. H. R. (2005). Common and distinct brain activation to threat and safety signals in social phobia. *Neuropsychobiology*, *52*(3), 163–168.

Strube, M. J. (2012). From "out there" to "in here": Implications of self-evaluation motives for self-knowledge. In S. V. &. T. Wilson (Ed.), *Handbook of self-knowledge* (pp. 397–412). The Guilford Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning, second edition: An Introduction*. MIT Press.

Swann, W. B. (1983). Self-verification: Bringing social reality into harmony with the self. In *Psychological Perspectives on the Self, Vol. 2* (pp. 33–66). https://doi.org/10.1126/science.218.4574.782

Swann, W. B., De La Ronde, C., & Hixon, J. G. (1994). Authenticity and positivity strivings in marriage and courtship. *Journal of Personality and Social Psychology*, *66*(5), 857–869.

Swann, W. B., Jr. (2012). Self-Verification Theory. In P. A. M. Van Lange Arie W. Kruglanski E. Tory Higgins (Ed.), *Handbook of theories of social psychology* (Vol. 2, pp. 23–42). SAGE Publications.

Swann, W. B., & Read, S. J. (1981a). Acquiring self-knowledge: The search for feedback that fits. *Journal of Personality and Social Psychology*, *41*(6), 1119–1128.

Swann, W. B., & Read, S. J. (1981b). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, *17*(4), 351–372.

Tangney, J. P., Miller, R. S., Flicker, L., & Barlow, D. H. (1996). Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, *70*(6), 1256–1269.

Taylor, S. E., & Brown, J. D. (1988). Illusion and Well-Being: A Social Psychological Perspective on Mental Health. *Psychological Bulletin*, *103*(2), 193–210.

Taylor, S. E., Neter, E., & Wayment, H. A. (1995). Self-evaluation processes. *Personality & Social Psychology Bulletin*, *21*(12), 1278–1287.

Tesser, A., & Cornell, D. P. (1991). On the confluence of self processes. *Journal of Experimental Social Psychology*, *27*(6), 501–526.

Tesser, A., Millar, M., & Moore, J. (1988). Some affective consequences of social comparison and reflection processes: The pain and pleasure of being close. *Journal of Personality and Social Psychology*, *54*(1), 49–61.

Tillfors, M., Furmark, T., Marteinsdottir, I., Fischer, H., Pissiota, A., Långström, B., & Fredrikson, M. (2001). Cerebral blood flow in subjects with social phobia during stressful speaking tasks: a PET study. *The American Journal of Psychiatry*, *158*(8), 1220–1226.

Touroutoglou, A., Hollenbeck, M., Dickerson, B. C., & Feldman Barrett, L. (2012). Dissociable large-scale networks anchored in the right anterior insula subserve affective experience and attention. *NeuroImage*, *60*(4), 1947–1958.

Tracy, J. L., & Robins, R. W. (2004). Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry*, *15*(2), 103–125.

Tracy, J. L., & Robins, R. W. (2007a). Emerging insights into the nature and function of pride. *Current Directions in Psychological Science*, *16*(3), 147–150.

Tracy, J. L., & Robins, R. W. (2007b). The psychological structure of pride: a tale of two facets. *Journal of Personality and Social Psychology*, *92*(3), 506–525.

Ubl, B., Kuehner, C., Kirsch, P., Ruttorf, M., Diener, C., & Flor, H. (2014). Altered neural reward and loss processing and prediction error signalling in depression. *Social Cognitive and Affective Neuroscience*, *10*(8), 1102–1112.

Ullsperger, M., Harsay, H. A., Wessel, J. R., & Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure & Function*, *214*(5–6), 629–643.

Villano, W. J., Kraus, N. I., Reneau, T. R., Jaso, B. A., Otto, A. R., & Heller, A. S. (2023). Individual differences in naturalistic learning link negative emotionality to the development of anxiety. *Science Advances*, *9*(1), eadd2976.

Vroling, M. S., & de Jong, P. J. (2009). Deductive Reasoning and Social Anxiety: Evidence for a Fear-confirming Belief Bias. *Cognitive Therapy and Research*, *33*(6), 633–644.

Wager, T. D., & Barrett, L. F. (2017). From affect to control: Functional specialization of the insula in motivation and regulation. In *bioRxiv* (p. 102368). https://doi.org/10.1101/102368

Watson, J. B., & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology*. https://psycnet.apa.org/journals/xge/3/1/1/

Wells, A. (2011). *Metacognitive Therapy for Anxiety and Depression*. Guilford Press.

Wells, A., Clark, D. M., Salkovskis, P., Ludgate, J., Hackmann, A., & Gelder, M. (1995). Social phobia: The role of in-situation safety behaviors in maintaining anxiety and negative beliefs. *Behavior Therapy*, *26*(1), 153–161.

Will, G.-J., Moutoussis, M., Womack, P. M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., NSPN Consortium, Rutledge, R. B., & Dolan, R. J. (2020). Neurocomputational mechanisms underpinning aberrant social learning in young adults with low self-esteem. *Translational Psychiatry*, *10*(1), 96.

Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *ELife*, *6*, 1–21.

Williams, L. A., & DeSteno, D. (2008). Pride and perseverance: the motivational role of pride. *Journal of Personality and Social Psychology*, *94*(6), 1007–1017.

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin*, *90*(2), 245–271.

Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. F. S. (2016). Self-other mergence in the frontal cortex during cooperation and competition. *Neuron*, *91*(2), 482–493.

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin*, *106*(2), 231–248.

Yoon, L., Somerville, L. H., & Kim, H. (2018). Development of MPFC function mediates shifts in self-protective behavior provoked by social feedback. *Nature Communications*, *9*(1), 3086.

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2003). *Schema therapy: A practitioner's guide*. Guilford Publications.

Zald, D. H., Boileau, I., El-Dearedy, W., Gunn, R., McGlone, F., Dichter, G. S., & Dagher, A. (2004). Dopamine transmission in the human striatum during monetary reward tasks. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *24*(17), 4105–4112.

Zamfir, E., & Dayan, P. (2022). Interactions between attributions and beliefs at trial-by-trial level: Evidence from a novel computer game task. *PLoS Computational Biology*, *18*(9), e1009920.

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, *146*(2), 118–149.

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707.